Proceedings of the Federated Conference on Computer Science and Information Systems pp. 477–482
DOI: 10.15439/2016F153
ACSIS, Vol. 8. ISSN 2300-5963

First Automatic Fongbe Continuous Speech Recognition System: Development of Acoustic Models and Language Models

Fréjus A. A. LAleye*‡, Laurent Besacier†, Eugène C. Ezin‡ and Cina Motamed*
*Laboratoire d'Informatique Signal et Image de la Côte d'Opale
Université du Littoral Côte d'Opale, France.
50 rue F. Buisson
BP 719, 62228 Calais Cedex
Email: {laleye, motamed}@lisic.univ-littoral.fr
†Laboratoire LIG- Univ. Grenoble Alpes - BP 53,
Email: laurent.besacier@imag.fr
‡Unité de Recherche en Informatique et Sciences Appliquées
Institut de Mathématiques et de Sciences Physiques
Université d'Abomey-Calavi Bénin,
BP 613 Porto-Novo.
Email: {frejus.laleye, eugene.ezin}@imsp-uac.org

***Abstract*—This paper reports our efforts toward an ASR system for a new under-resourced language (Fongbe). The aim of this work is to build acoustic models and language models for continuous speech decoding in Fongbe. The problem encountered with Fongbe (an African language spoken especially in Benin, Togo, and Nigeria) is that it does not have any language resources for an ASR system. As part of this work, we have first collected Fongbe text and speech corpora that are described in the following sections. Acoustic modeling has been worked out at a graphemic level and language modeling has provided two language models for performance comparison purposes. We also performed a vowel simplification by removing tones diacritics in order to investigate their impact on the language models.**

I. Introduction

AUTOMATIC Speech Recognition (ASR) is a technology that allows a computer to identify the words spoken by a person in microphone. Speech recognition technology is changing the way information is accessed, tasks are accomplished and business is done. The growth of speech applications over the past years has been remarkable [1]. ASR applications have been successfully achieved for most western languages such as English, French, Italian etc., for Asian languages such as Chinese, Japanese, Indian etc, because of the large quantity and the availability of linguistic resources of these languages [2]. This technology is less prevalent in Africa despite its 2,000 languages because of lack or unavailability of these resources for most African languages (vernacular for most). Also, for the most of the time, these are not written languages (no formal grammar, limited number of dictionaries, few linguists). Despite the shortcomings, some have been investigated and now have the linguistic resources to build a speech recognition systems. For example, in the context of a project entitled ALFFA[1], the authors in [3] developed ASR systems for 4 sub-saharan african languages (Swahili, Hausa, Amharic and Wolof). Another language of West Africa (Yoruba) spoken mainly in Nigeria, in Benin and neighboring countries has been also investigated for an ASR system. [4] provides a brief review of research progress on Yorúbà Automatic Speech Recognition.

Our main objective in this paper is to introduce a first ASR system for an under-resourced language, Fongbe. Fongbe is a vernacular language spoken primarily in Benin, by more than 50% of the population, in Togo and in Nigeria. It's an under-resourced because it lacks linguistics resources (speech corpus and text data) and very few websites provide textual data. Building these resources, acoustic models and language models for Fongbe ASR becomes a challenging task. For this, we used Kaldi toolkit[2] that has allowed us to train our acoustic models on speech data that we have collected ourselves. For the language modeling, we used SRILM toolkit[3] to built trigram language models that we trained on collected text data. To enhance performance of our ASR, we subsequently transformed the vowels by normalizing different tones of Fongbe. Experiments have shown a significant improvement in the results given by the world error rates (WER).

The remainder of this paper is organized as follows. The next section describes the target under-resourced language that is Fongbe. Section 3 describes how text and speech corpora have been collected. Section 4 and 5 focus respectively on language modeling and acoustic modeling. Section 6 presents

[1] http://alffa.imag.fr
[2] kaldi.sourceforge.net/
[3] www.speech.sri.com/projects/srilm/

978-83-60810-90-3/$25.00©2016, IEEE

and comments the experimental results of WER that we obtained. Section 7 concludes this paper and presents future work.

II. Description of Fongbe language

Fongbe language is the majority language of Benin, which is spoken by more than 50% of Benin's population, including 8 million speakers and also spoken in Nigeria and Togo. The Fongbe people are the largest ethnic group in Benin. Fongbe is part of the Gbe dialect cluster and is spoken mainly in Benin [6]. It is quite widespread in the media and is used in schools, including adult literacy. The Fongbe group is one of the five Gbe dialect. J. Greenberg classifies Fongbe in the Kwa languages group in the Niger-Congo branch of the large family Niger-Kordofan [5]. It is written officially in Benin with an alphabet derived from the Latin writting since 1975. It has a complex tonal system, with two lexical tones, high and low, which may be modified by means of tonal processes to drive three further phonetic tones: rising low-high, falling high-low and mid [6]. The use of diacritical marks to transcribe the different tones of the language is essential even if they are not always marked since Fongbe is originally a spoken language. The Fongbe's vowel system is well suited to the vocalic timbre as it was designed by the first Phoneticians. It includes twelve timbres: 7 oral vowels with 4 degrees of aperture and 5 nasal vowels with 3 degrees of aperture. Its consonant system includes 22 phonemes.

Scientific studies on the Fongbe started in 1963 with the publication of Fongbe-French dictionary [7]. Since 1976, several linguists have worked on the language and many papers were published on the linguistic aspects of Fongbe. Unlike most of the western languages (English, French, Spanish, etc) and some Asian languages (Chinese, Japanese, etc) and African (Wolof, Swahili, shrugged, etc.) the Fongbe language suffers from a very significant lack of linguistic resources in digital form (text corpus and speech) despite the many linguistic works (phonology, lexicon and syntax).

III. Collection of language resources

The development of automatic continuous speech recognition system is made from a large amount of data which must contain both speech signals (for the acoustic modeling of the system) and also text data (for the language model of the system). It becomes a challenge and very difficult when it is an under-resourced language that still doesn't possess these digital resources. In this section we describe the methodology used to collect texts and audio signals of Fongbe language for building of the recognition system.

A. Speech corpus

As an audio corpus is not available for Fongbe, we proceeded to the speech signals collection to build the audio data for the system. We thus conducted the tedious task of recording the texts pronounced by native speakers (including 8 women and 20 men) of Fongbe in a noiseless environment. We have recorded at 16Khz 28 native speakers who have spoken around 1500 phrases (from daily living) grouped into 3 categories. A category is read by several speakers and contains texts that are different from contents of other categories. These recordings were made with an android application referred to as LigAikuma [8] which is developed by GETALP group of Grenoble's Computer Science Laboratory. Overall, there are around 10 hours of speech data that have been collected. First, we split the data by categories leading to a first configuration FC1: 2 categories for training (8 hours) and 1 category for testing (2 hours). Next, we split the data by speakers leading to a second configuration FC2: 20 speakers (8 hours) for training and 8 speakers (2 hours) for testing. We split the data this way firstly to make sure that category appear in test data will not appear in training and secondly, to reduce the chance of having speakers overlapping between training and testing.

TABLE I
Contents of Fongbe speech corpus.

	Speech segments	Phrases	Duration	Categories	Speakers
		FC1 - config			
Train data	8,234	879	7h 35mn	C2 & C3	25
Test data	2,168	542	1h 45mn	C1	4
		FC2 - config			
Train data	8,651	1,421	8h	C1, C2 & C3	21
Test data	1,751	1,410	2h	C1, C2 & C3	7

B. Text corpus

To build a language model we need to have a text corpus containing thousands of words of the given language. The standard way most commonly used to build a text corpus is the collection of texts from websites. As we have shown in previous sections, Fongbe is an under-resourced language and thus has a very limited number of websites compared to languages such as Wolof, Hausa, and above all Arabic, French and English that have a very large wide coverage on the internet and do not suffer from lack of textual data. So, based on the few websites that provide texts in Fongbe, we used RLAT [9] to crawl text from these websites covering few texts from everyday life and many texts of the Bible translated into Fongbe. RLAT enables us to crawl text from a given webpage with different link depths. For improving the quantity of texts obtained from HTML links of websites, we have added to our corpus some texts obtained from PDF files that cover many of Fongbe citations, songs and the Universal Declaration of Human Rights. After extracting all text content in web pages and pdf file, we conducted to a cleaning and normalization of the texts:

1) remove all HTML tags and codes,
2) remove empty lines and punctuations,
3) conversion of texts to Unicode,
4) remove pages and lines from other languages than Fongbe,

5) transcription of special characters and numbers,
6) delete duplicate lines.

In total, we obtained nearly 10,130 words to build our vocabulary dictionary and a corpus which contains 34,653 sentences collected from the few documents written in Fongbe that are actually available. In table II, we list the websites used to extract text for two language models (LM1 and LM2) and from which we selected 1,500 utterances (source 1) for recording speech data for the training and testing set.

TABLE II
CONTENTS OF TEXT CORPUS.

Source	Websites	Text	utterances
1	http://www.fonbe.fr	variety of texts in daily life	1,500
2	http://unicode.org/udhr/d/udhr_fon.txt	Universal Declaration of Human Rights	92
3	http://ipedef-fongbe.org/	Educational texts, songs and tales	2,200
4	http://www.vodoo-beninbrazil.org/fon.html	Educational Texts	1,055
5	https://www.bible.com/fr/bible/813/dan	The Bible	29,806

IV. LANGUAGE MODELING

Statistical language models (LM) are employed in various natural language processing applications, such as machine translation, information retrieval or automatic speech recognition. they describe relations between words (or other tokens), thus enabling to choose most probable sequences. This proves to be especially useful in speech recognition, where acoustical models usually produce a number of hypotheses, and re-ranking them according to a language model can substantially improve recognition rates [10] To compare the performance of our Fongbe recognition system, we built two language models (LM) using the same text corpus. The first language model (LM1) is built with the original texts after normalization and contain different tonal vowels. The use of tonal vowels implies that the system has to handle 26 vowels (with accented characters) considered as different tones instead of the 12 initial vowels. The second language model is built with the original texts that we modified by performing a second normalization on different tonal vowels from text corpus. The normalization was made by removing the tones from vowels and replacing accented characters by single characters. The result is that we have new entries with their transcriptions in our vocabulary dictionary. For example, the original word *axɔ́sú*, which means *king* will become in the dictionary *axɔsu*. Table III summarizes the various changes made to the vowels.

We used SRILM toolkit to train the two languages models. LM1 and LM2 were trained on 995,338 words (10,095 unigrams) by using the training data from text corpus (1,054,724 words, 33,153 sentences) without utterances used for the speech corpus (5,490 words and 1,500 sentences removed). LM1 was trained with the original texts while LM2 was trained with the modified texts by vowel normalization. To represent the uncertainty of our language models, we calculate the perplexity values of all the utterance transcriptions from speech corpus that are not contained in the various text corpus and which represents our test data to evaluate the performance of the two language models. Table V shows the perplexity values. The vowel normalization after the original text modification has positive impact on the quality of the language model by reducing in the OOV from 9.1% to 4.96%. This leads to observe a significant perplexity improvement with LM2 compared to LM1. Final system has been built using a lexicon which contains 10,130 unique grapheme words. As in [12], [11], we used grapheme as modeling unit to create our own lexicon because. An example of its content obtained after text pre-processing is shown in Table IV.

TABLE III
VOWEL NORMALIZATION.

Tonal vowels	Normalization
á	/a/
à	/a/
ã	/a/
ó	/o/
ò	/o/
õ	/o/
é	/e/
è	/e/
ẽ	/e/
ú	/u/
ù	/u/
ũ	/u/
í	/i/
ì	/i/
ĩ	/i/
ɛ́	/ɛ/
ɛ̀	/ɛ/
ɛ̃	/ɛ/
ɔ́	/ɔ/
ɔ̀	/ɔ/
ɔ̃	/ɔ/

TABLE IV
EXAMPLE OF LEXICON'S CONTENT

	Word	Graphemes
Original text	axɔ́súɖuɖu hãgbɛ́	a x ɔ́ s ú ɖ u ɖ u h ã g b ɛ́
Vowel normalization	axɔsuɖuɖu haagbɛ	a x ɔ s u ɖ u ɖ u h a a g b ɛ

TABLE V
LANGUAGE MODEL COMPARISON USING THE PERPLEXITY.

LM	Vocab (words)	OOV	PPL
LM1	10,130	9.1%	591
LM2	8,244	4.96%	138

V. ACOUSTIC MODELING

In this section, we describe the methods that we used for training and testing our 2 configurations (FC1 and FC2) and present in the next section the obtained results. The recordings and their transcriptions are used for acoustic modeling. The Acoustics models (AMs) are trained and tested on acoustic data from both FC1 and FC2 by using Kaldi acoustic modeling scripts that we have adapted to produce Kaldi scripts for Fongbe. We not only explored AM training methods but also experimented the impact of presence of tones in the utterances transcription from speech corpus by using LM1 (with tones) or LM2 (no tones). Thus, FC1 and FC2 training are performed not only with the same scripts but also by using both pronunciation dictionary. The pronunciation dictionary based grapheme that is used with LM1 contains 49 graphemes while the dictionary used with LM2 contains 28 graphemes.

The models are trained with 13 MFCC (Mel-Frequency Cesptral Coefficients) features whose coefficients are tripled with the $\Delta+\Delta\Delta$ by computing the first and second derivatives from MFCC coefficients. We also computed other feature transformation techniques such as LDA (Linear Discriminant Analysis) and MLLT (Maximum Likelihood Linear Transform) which gain substantial improvement over $\Delta+\Delta\Delta$ transformation. Subsequently, we also applied speaker Adaptation with feature-space Maximum Likelihood Linear Regression (fMLLR). Refer to the papers [13] and [14] for details on the theory of these transformation techniques implemented in Kaldi ASR. Figure 1 and Table VI show the hierarchy of the acoustics models that we trained in our experiments. In this hierarchy, we started by training monophone model using the MFCC features and we ended up training of SGMM using fMMI transformed features. The intermediate triphone models are also trained as shown in Figure 1. For decoding, we used the different trained acoustics models with the utterances from the test data. For each trained acoustic model we used the same speech parametrization and feature transformation method as was used for the given acoustic model at training time.

TABLE VI
ACOUSTICS MODELS. COMBINE* REPLACED COMBINE_TRI3B_FMMI_INDIRECT_SGMM2_5B2_MMI_B0.1

Training method	Script
Monophone	mono
Triphone	tri1
$\Delta+\Delta\Delta$	tri2a
LDA + MLLT	tri2b
LDA + MLLT + SAT + FMLLR	tri3b
LDA + MLLT + SAT + FMLLR + fMMI	tri3b_fmmi_a
LDA + MLLT + SAT + FMLLR + MMI	tri3b_mmi_b0.1
LDA + MLLT + SAT + FMLLR + fMMI + MMI	tri3b_fmmi_indirect
LDA + MLLT + SGMM	sgmm2_5b2
LDA + MLLT + SGMM + MMI	sgmm2_5b2_mmi_b0.1
LDA + MLLT + SGMM + fMMI + MMI	combine*

VI. EXPERIMENTAL RESULTS

The experiments focus on comparing the quality of ASR hypothesis measured by WER on AMs trained by different methods. To obtain the best path, we followed the standard KALDI procedures and report the best WER. The experiments were performed first on LM1 built with the original texts and using both speech data configurations. Then we conducted experiments based on the same procedures on LM2 including texts without diacritics. The interest is to measure the impact of using diacritics in language modelling from the results given by the WER. We also showed how the data speech configuration influence the quality of AMs measured by WER.

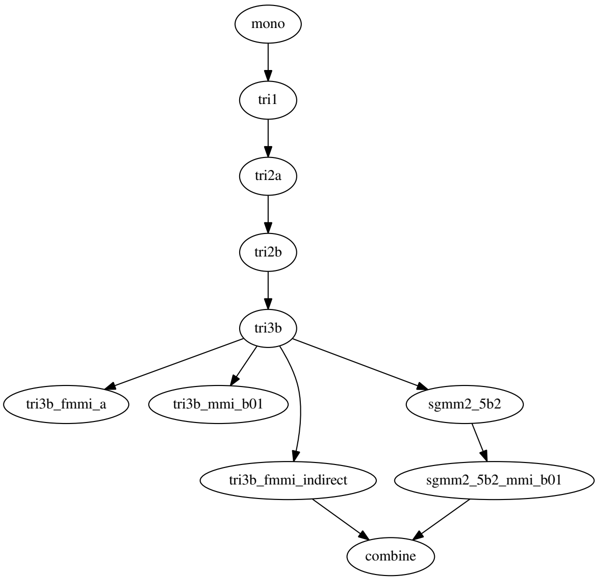

Fig. 1. Hierarchy of trained 2coustics models

A. Results before vowel normalization

In this subsection we present the results of different acoustic training methods according to data speech configuration. Table VII presents AMs results for LM1.

From the results in table VII, we can see that the monophone AM has the worst WER while the best performances are achieved with the sgmm2_5b2 AM for FC1-config and the sgmm2_5b2_mmi for FC2-config. We can thus notice that the monophone AM is typically used for the initialization of triphone models. The quality of speech recognition varies according to the used discriminative training method. The LDA+MLLT is more effective feature transformation than using $\Delta + \Delta\Delta$ features. There are subtle performance differences among the discriminatively trained acoustic model. The WER on both speech data configuration for fixed LM1 is around 44%. This can be explained by the complexity of Fongbe language for modelling the diacritics and the quality of language model used (LM1). The perplexity reported in table V justifies this assertion. Figure VI-A shows the curve performances of the acoustic training methods for both speech data configuration.

B. Results after vowel normalization

Table VIII presents two WERs of different acoustic training methods according to data speech configuration. In the second

TABLE VII
WER OF LM1-BASED ASR (WITH DIACRITICS) FOR DIFFERENT TRAINING MONOPHONE AND TRIPHONE METHODS

Speech data config/ method	WER %
FC1-config	
Monophone (a)	69.44
Triphone (b)	69.13
$\Delta + \Delta\Delta$ (c)	70.21
LDA + MLLT (d)	65.7
LDA + MLLT + SAT + FMLLR (e)	54.96
LDA + MLLT + SAT + FMLLR + fMMI (f)	55.36
LDA + MLLT + SAT + FMLLR + MMI (g)	51.11
LDA + MLLT + SAT + FMLLR + fMMI + MMI (h)	55.60
LDA + MLLT + SGMM (i)	**44.04**
LDA + MLLT + SGMM + MMI (j)	47.11
LDA + MLLT + SGMM + fMMI + MMI (k)	49.83
FC2-config	
Monophone (a)	71.97
Triphone (b)	60.37
$\Delta + \Delta\Delta$ (c)	59.74
LDA + MLLT (d)	57.52
LDA + MLLT + SAT + FMLLR (e)	51.47
LDA + MLLT + SAT + FMLLR + fMMI (f)	53.06
LDA + MLLT + SAT + FMLLR + MMI (g)	52.75
LDA + MLLT + SAT + FMLLR + fMMI + MMI (h)	52.37
LDA + MLLT + SGMM (i)	49.85
LDA + MLLT + SGMM + MMI (j)	**44.09**
LDA + MLLT + SGMM + fMMI + MMI (k)	44.17

TABLE VIII
WER OF LM2-BASED ASR (WITHOUT DIACRITICS) AND LM1'-BASED ASR (REMOVING OF DIACRITICS FROM HYPOTHESES AND REFERENCES OF LM1-BASED ASR).

Speech data config/ method	LM2-Based ASR	LM1'-Based ASR
FC1-config		
Monophone (a)	36.36	59.05
Triphone (b)	28.19	46.8
$\Delta + \Delta\Delta$ (c)	28.21	46.98
LDA + MLLT (d)	24.4	41.52
LDA + MLLT + SAT + FMLLR (e)	17.83	29.29
LDA + MLLT + SAT + FMLLR + fMMI (f)	19.72	31.34
LDA + MLLT + SAT + FMLLR + MMI (g)	18.93	35.59
LDA + MLLT + SAT + FMLLR + fMMI + MMI (h)	18.26	35.44
LDA + MLLT + SGMM (i)	15.23	**20.56**
LDA + MLLT + SGMM + MMI (j)	15.3	20.68
LDA + MLLT + SGMM + fMMI + MMI (k)	**14.83**	21.39
FC2-config		
Monophone (a)	52.26	57.89
Triphone (b)	38.72	47.47
$\Delta + \Delta\Delta$ (c)	38.58	46.39
LDA + MLLT (d)	35.34	42.45
LDA + MLLT + SAT + FMLLR (e)	30.74	35.63
LDA + MLLT + SAT + FMLLR + fMMI (f)	35.36	37.46
LDA + MLLT + SAT + FMLLR + MMI (g)	32.38	36.19
LDA + MLLT + SAT + FMLLR + fMMI + MMI (h)	32.94	37.52
LDA + MLLT + SGMM (i)	31.64	**31.58**
LDA + MLLT + SGMM + MMI (j)	31.36	32.75
LDA + MLLT + SGMM + fMMI + MMI (k)	**28.93**	32.02

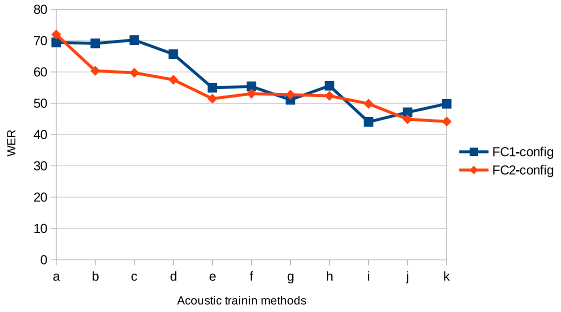

Fig. 2. Influence of speech data configuration on speech recognition quality. LM2 is fixed and only speech data and acoustic models vary. The letter in abscissa represent acoustic training methods labelled in table VI-A

column (LM2-Based ASR), we have included the WER results of ASR performed after vowel normalization (without diacritics).

Colunm of LM2-Based ASR in Table VIII also shows that triphone models significantly improve the monophone model performance. The tri2b+SAT+FMLLR acoustic model adapted to speaker from feature-space Maximum Likelihood Linear Regression reduced the WER by 6% absolute for both speech data configuration. The WER on FC1-config is lower than 20% for discriminative methods based on tri3b. For FC2-config, these acoustic training methods reduced the WER by 20%. The best results are coming from the training for Subspace Gaussian Mixture Models (SGMM), with an overall WER of 14.83% for FC1-config and 28.93% for FC2-config. The speech data divided by speakers helps us to obtain a relative gain of 14% with the best final WER of 14.83%. This leads us to choose AM training methods using SGMM for performance comparison among FC1-config and FC2-config. Figure VI-B shows the evolution of WER depending on acoustic models with LM2.

It is therefore remarkable that the language model LM2 gives very satisfactory decoding results compared to LM1 standard (with diacritics). Adding diacritics in text corpus before language modelling maked the speech recognition system less efficient by increasing the WER by 44.04% compared to 15.23% (performance without diacritics). While diacritics add information, which should help the recognition system, it also increases OOV rate and perplexity of the language model (see table V).

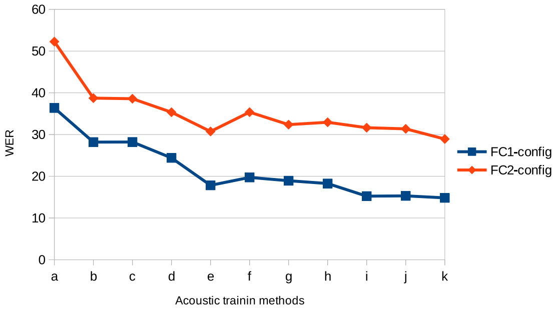

Fig. 3. Influence of speech data configuration on speech recognition quality. LM2 is fixed and only speech data and acoustic models vary. The letter in abscissa represent acoustic training methods labelled in table VI-B

For further, we performed an effective comparison of ASR performance without diacritics. To do this, we removed the diacritics from the outputs (hypotheses) and references of ASR system built with LM1 (LM1'-Based ASR). The obtained results for this evaluation are included in the third column of Table VIII. These results can be compared to results obtained with LM2-Based ASR system (second column). This comparison leads us to assert that the removing of diacritics for different models is more effective and provides an efficient ASR system.

VII. Conclusion

In this work we introduced the first system of Fongbe continuous speech recognition by training different acoustic models using Kaldi scripts and different language models using SRILM toolkit. We also demonstrated the effect of tones on the quality of the recognition system. This leads us to conclude that with the current state of our system, the language modelling without diacritics improves significantly the recognition performances by decreasing the WER by 15.23% for speech data divided by speakers and 28.93% for speech data divided by category. Using the Kaldi recipe and the language resources we provide, researcher can build a Fongbe recognition system with the same WER obtained in this paper. For future work, firstly we will enhance the speech and text data and introduce other training techniques to further improve the performance of this first system of Fongbe recognition. Secondly, we will investigate the Fongbe re-diacritization in the context of Speech recognition.

References

[1] J. K. Tamgno and E. Barnard and C. Lishou and M. Richomme, *Wolof Speech Recognition Model of Digits and Limited-Vocabulary Based on HMM and ToolKit*, in. 14th International Conference on Computer Modelling and Simulation (UKSim), pp. 389–395, 2012 UKSim.

[2] Besacier, L., Barnard, E., Karpov, A., and Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

[3] E. Gauthier and L.Besacier and S. Voisin and M. Melese and U. P. Elingui *Collecting Resources in Sub-Saharan African Languages for Automatic Speech Recognition: a Case Study of Wolof* in. 10th edition of the Language Resources and Evaluation Conference, 23-28 May 2016, Slovenia.

[4] S. A. M. Yusof and A. F. Atanda and M. Hariharan, *A review of Yorúbà Automatic Speech Recognition*, in. System Engineering and Technology (ICSET), IEEE 3rd International Conference on, pp. 242–247, Aug.2013.

[5] J. Greenberg, *Languages of Africa*, La Haye Mouton, pp. 177, 1966.

[6] C. Lefebvre and A-M. Brousseau, *A grammar of Fonge*, De Gruyter Mouton, PP. 608, December 2001.

[7] A. B. AKOHA, *Syntaxe et lexicologie du Fon-gbe: Bénin*, Ed. L'harmattan, pp. 368, January 2010.

[8] Blachon, D., Gauthier, E., Besacier, L., Kouarata, G.-N., Adda-Decker, M., and Rialland, A. (2016). Parallel speech collection for under-resourced language studies using the LIG-Aikuma mobile device app. In *Proceedings of SLTU (Spoken Language Technologies for Under-Resourced Languages)*, Yogyakarta, Indonesia.

[9] A. W. Black and T. Schultz, *Rapid Language Adaptation Tools and Technologies for Multilingual Speech Processing*, in Automatic Speech Recognition & Understanding, IEEE Workshop, pp. 51, 2009.

[10] Sebastian Dziadzio, Aleksandra Nabozny, Aleksander Smywinski-Pohl and Bartosz Ziolko, *Comparison of Language Models Trained on Written Texts and Speech Transcripts in the Context of Automatic Speech Recognition*, in Proc. Proceedings of the IEEE Federated Conference on Computer Science and Information Systems, 5, pp. 193-197, Pologne 2015.

[11] S. Seng and S. Sam and V. Bac Le and B. Bigi and L. Besacier, *Which units for acoustic and language modeling for Khmer automatic speech recognition?*, SLTU 2008.

[12] J. Billa and all, *Audio indexing of Arabic broadcast news*, in Proc. IEEE International Conference on Acoustique, Speech and Signals Processing, pp. 5-8, Orlando 2002.

[13] D. Povey and A. Ghoshal et al., *The Kaldi Speech Recognition Toolkit*, in IEEE ASRU, 2011.

[14] D. Povey and G.Saon, *Feature and model space speaker adaptation with full covariance Gaussians*, in INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006

[15] Lukasz Laszko, *Word detection in recorded speech using textual queries*, in Proc. Proceedings of the IEEE Federated Conference on Computer Science and Information Systems, 5, pp. 849-853, Pologne 2015.