Speed perturbation and vowel duration modeling for ASR in Hausa and Wolof languages

Elodie Gauthier, Laurent Besacier, Sylvie Voisin

Speed perturbation and vowel duration modeling for ASR in Hausa and Wolof languages

Elodie Gauthier[1], *Laurent Besacier*[1], *Sylvie Voisin*[2]

[1]Laboratoire d'Informatique de Grenoble (LIG), Univ. Grenoble Alpes, Grenoble, France
[2]Laboratoire Dynamique Du Langage (DDL), CNRS - Université Aix Marseille, France

elodie.gauthier@imag.fr, laurent.besacier@imag.fr, sylvie.voisin@cnrs.fr

Abstract

Automatic Speech Recognition (ASR) for (under-resourced) Sub-Saharan African languages faces several challenges: small amount of transcribed speech, written language normalization issues, few text resources available for language modeling, as well as specific features (tones, morphology, etc.) that need to be taken into account seriously to optimize ASR performance. This paper tries to address some of the above challenges through the development of ASR systems for two Sub-Saharan African languages: Hausa and Wolof. First, we investigate data augmentation technique (through speed perturbation) to overcome the lack of resources. Secondly, the main contribution is our attempt to model vowel length contrast existing in both languages. For reproducible experiments, the ASR systems developed for Hausa and Wolof are made available to the research community on *github*. To our knowledge, the Wolof ASR system presented in this paper is the first large vocabulary continuous speech recognition system ever developed for this language.

1. Introduction

A major challenge of the digital market in Africa is to ensure that online applications and services become accessible to a wide range of people, be they literate or not. For the latter ones, applications should be able to talk and listen to Africans in the true sense of the words and several UNESCO reports make explicit reference to text to speech synthesis (TTS) and automatic speech recognition (ASR) as a technological facilitator. If we focus on automatic speech recognition (ASR), some challenges are related to the small amount of transcribed speech (a consequence being the lack of speaker diversity during training), written language normalization issues, few text resources available for language modeling, as well as specific features (tones, morphology) that need to be taken into account seriously to optimize ASR performance. More challenges are given in a survey on ASR for under-resourced languages published in [1].

Paper objective and contributions

This paper tries to address some of the above challenges through the development of ASR systems for two Sub-Saharan African languages: Hausa and Wolof. In term of phonology, their common characteristic is to appear with vowel length contrast. In other words, two versions (short / long) of the same vowel exist in the phoneme inventory of the language. We expect that taking into account this contrast in ASR might help and this is the first problem addressed in this paper. Also, since the transcribed speech databases available for these languages are limited, our second contribution is to investigate a data augmentation technique (through speed perturbation) to improve deep neural networks (DNN) based ASR. Finally, since for the Wolof language (mostly spoken and seldom written), few literate people can be found to correctly utter read speech, we also present a fast data cleaning procedure to improve the quality of the development and test set recorded. For reproducible experiments, the ASR systems developed for Hausa and Wolof are made available on our *github* repository[1]. It is worth mentioning that the Wolof ASR system presented in this paper is the first large vocabulary continuous speech recognition system ever developed for this language (only a digit and small vocabulary recognition system in Wolof was previously presented in [2]).

Languages studied

Hausa and Wolof are two African languages, largely spoken in the West of the continent. Hausa is part of the Afro-Asiatic phylum. The language is one of the most spoken languages in all the continent with more than 35 million of speakers (mostly in Niger and Nigeria) as a first language. Hausa is also widely used as a common language with an additional 15 million people using it as second and third language.

Wolof is part of the Niger-Congo phylum. The language is considered as one of the common languages of West Africa. It is mostly spoken in Senegal (40% as native language by the Wolof people), Gambia (by 12% of the Wolof people) and Mauritania (by 8% of the Wolof people). The Wolof spoken in Mauritania is the same as the one spoken in Senegal, because the minority of speakers is located around the Senegal river that they share. But Gambian Wolof is quite different and two ISO 639-3 language code exist to distinguish both Wolof (respectly "WOL" and "WOF"). For our studies, we focus on the Senegalese Wolof because of population coverage.

Hausa and Wolof both have length contrast at the phone level. Nonetheless, this phenomenon does not appear in the same way in both languages: contrast can be located on vowels as well as on consonants in Wolof while in Hausa this feature is only for vowels. Both languages have different complex vocalic system. The Wolof system attests 8 short vowels /i/, /e/, /ɛ/, /a/, /ə/, /ɔ/, /o/ and /u/. Each of them, except /ə/, have a corresponding long phoneme. The Hausa presents a vocalic system with 5 vowels for which each one can be opposed to a corresponding long vowel. There are also 2 diphtongs /ai/ and /au/. Concerning the tones, Wolof is a non-tonal language unlike the majority of Sub-Saharan Africa languages. In Hausa, there are three tones: high, low and falling tones [3]. Length contrast and tones change the meaning of a word, but tones are not considered yet in this study.

Paper outline

This paper is organized as following. In Section 2, we summarize the recent works on data augmentation for ASR as well as on vowel length contrast modeling in ASR. In Section 3, we describe the data used and the baseline ASR systems de-

[1]https://github.com/besacier/ALFFA_PUBLIC/tree/master/ASR

signed. Then, Section 4 details our speed perturbation experiments applied to Hausa and Wolof ASR. Section 5 illustrates the vowel length contrast in both languages and shows that length-contrasted DNN models are better than non length-contrasted models for ASR. Finally, Section 6 concludes this paper and gives a few perspectives.

2. Related Works

2.1. Data augmentation for ASR

Data augmentation consists of artificially increasing the quantity of training data and has been widely used in image processing [4] and speech processing against noise [5]. Recently, vocal tract length perturbation (VTLP) [6] was applied successfully for deep neural networks based ASR. VTLP was also successful to improve low-resource ASR as shown in [7] for Assamese and Haitian Creole, two languages of the IARPA Babel program. Last year, [8] presented an audio augmentation technique with low implementation cost. This method, based on speed perturbation, lead to a better word error rate (WER) improvement than VTLP on multiple speech databases while being very simple to implement. However, in [8], speed perturbation was only applied to English and Mandarin (well-resourced) languages. So, one contribution of this paper is to evaluate speed perturbation with low resource scenarios (Hausa and Wolof ASR).

2.2. Phone duration modeling for ASR

As far as phone duration modeling in speech recognition is concerned, [9] modeled word duration at the acoustic modeling level. Speech recognition performance was improved by rescoring N-best lists with the duration models built. Few years later, [10] worked on duration modeling at both word and phone levels. Lattice rescoring was used but the WER reduction was limited. Nonetheless, the technique implemented improved the transcription quality when combining systems. [11] also focused on the phone length in Finnish and compared different duration modeling techniques. A 8% relative reduction of the letter error rate compared to the baseline system (not handling duration) was obtained. However, the decoding speed was reduced. Finally, [12] used phone duration modeling in Estonian. Three different phone lengths exist in Estonian (short, long and overlong). A method was presented based on linguistic and phonological characteristics of phone as well as their surrounding phones context to calculate a decision tree that classifies phones into groups of similar durations. ASR performance was, again, only slightly improved. All these approaches take for granted the phoneme duration contrast while we believe that for under-documented languages, it is important to empirically verify its concrete realization before deciding to separate a given phone in two short/long units for ASR models. Our methodology is explained in more details in Section 5.

3. Data collection and baseline ASR systems

3.1. Speech and text data used

3.1.1. The Globalphone Hausa corpus

We exploited the GlobalPhone Hausa Speech Corpus [13] to train our ASR system in Hausa and all the resources provided. It was recorded in Cameroon with native Hausa speakers who read utterances written with the *boko* orthography. The training set is composed of 6h36mins of speech data (5,863 utterances in total, read by 24 males and 58 females). For the decoding, the development set (*dev*) is composed of 1,021 utterances (a total speech duration of 1h02mins) read by 4 males and 6 females. For the evaluation, the test set (*test*) is composed of 1,011 utterances (a total speech duration of 1h06mins) read by 5 males and 5 females.

3.1.2. Speech and text data collected for Wolof

We collected our own data to build an ASR system for Wolof. We exploit a set of very few electronic documents available for this language, gathered as part of [14] to build our read speech corpus. These resources contain dictionary's sentences of [15], stories, proverbs, debates, lyrics. The read speech corpus was recorded in Dakar (Senegal) with native Wolof speakers (10 males, 8 females - from 24 to 48 years old). We recorded 21h22mins of speech signal (1,000 utterances read by each speaker). We split our collected speech corpus into 3 sets : the data set used to train the ASR system represents 16h49mins (8 males, 6 females have read 13,998 utterances in total); the development set (*dev*) contains 2h12mins of speech data (1 male, 1 female read 1,000 utterances each); the *test* set is composed of 1 male and 1 female, representing 2h20mins of speech (2,000 utterances read). Later, in order to build a language model with more textual data written in Wolof, we collected on the Web the Universal Declaration of Human Rights, the Bible and a book written by a humanist. We also crawled the Wikipedia database in Wolof using Wikipedia Extractor [16]. More description about the data collected in Wolof is available in [17].

3.1.3. Specific challenge for Wolof language: data cleaning

Our read speech corpus has been recorded with native Wolof speakers. However, Wolof is not learnt at school and our speakers faced reading issues. As a consequence, many speech utterances contain reading errors and differ from the initial transcript. While flawed transcripts are a well known problem for ASR training, we decided to focus, first, on cleaning the transcripts used for evaluation. Both *dev* and *test* sets were verified by an expert of Wolof. The task was to listen to the recordings and to confirm or not the perfect match between transcripts and audio. For this, we added a specific *check* mode to Lig-Aikuma[2], our mobile app recently developped for speech data collection [18]. It turned out that half of the sentences were not correctly read[3]. After this manual check, cleaned data sets for Wolof are the following (similar in size to Hausa *dev*/*test*):

- *dev* : 1,120 utterances (instead of 2,000) - duration is 1h12mn
- *test* : 846 utterances (instead of 2,000) - duration is 55mn

3.2. Baseline ASR systems for Hausa and Wolof

We used Kaldi speech recognition toolkit [19] for building our ASR systems for both languages. Our systems are CD-DNN-HMM hybrid systems. We trained the network using state-level minimum Bayes risk [20] (sMBR). The initial acoustic models were built using 13 Mel-frequency cepstrum coefficients (MFCCs) and Gaussian mixture models (GMMs) on 6.6h training data for Hausa and on 16.8h for Wolof. We trained triphone models by employing 2,887 context-dependent states for Hausa and 3,401 context-dependent states for Wolof, and 40K Gaussians for both. Then, the acoustic states were used as targets of the DNN (6 hidden layers with 1024 units each). The initial weights for the network were obtained using Restricted Boltzmann Machines (RBMs) [21] and fine tuning was done using Stochastic Gradient Descent. The standard Kaldi recipe for CD-DNN-HMM training was used (see [17] for more details).

[2] https://forge.imag.fr/frs/download.php/706/MainActivity.apk

[3] We can expect the same trend on training data but - since utterances checking is very time consuming - we did not apply it yet to our *train*. Moreover, training is rather robust to imperfect transcripts so we kept all the data to learn our ASR models.

Concerning the language model (LM), we used a trigram model, in the decoding process, for each Hausa and Wolof ASR systems. For Hausa, we used the language model provided by [13]. It was built from a text corpus of about 8M words. Its perplexity, calculated on the *dev* set and the *test* set is low: respectively, 88 (0.19% of out of vocabulary words (OOVs) and 90 (0.21% of OOVs)). For Wolof, we built our own statistical language model with the SRILM toolkit. We interpolated a language model built from two LMs: the first model was built from the very few initial electronic documents mentioned in 3.1.2 and the second from Web data we crawled. Finally, this interpolated language model of Wolof was built from a rather small text corpus of 615,631 words. On both cleaned *dev* and *test* sets, the perplexity is respectively 268 (4.2% of OOVs) and 273 (3.6% of OOVs).

About the pronunciation dictionary, we also used the one of [13] for the training and decoding stages of the Hausa ASR. It contains 38,915 entries and 33 phonemes. For Wolof, we used our in-house pronunciation dictionary which contains 32,039 entries and a phoneme inventory of 34 phonemes. As a seed, we used the entries of Diouf's [15] and Fal's [22] dictionaries for which the phonetic transcription were specified. Then, we trained a Grapheme-to-Phoneme (G2P) model to automatically transcribe into phonetic symbols the vocabulary of our text corpus (used to built the LM) not phonetized yet[4]. At this stage, phoneme inventories (for both languages) do not take into account any length contrast.

We can see in table 1 below, the performance for the first Hausa and Wolof CD-DNN-HMM ASR systems. For Wolof, we report results on both initial and cleaned data sets (see 3.1.3). It is important to mention first that, while being trained on limited data, both CD-DNN-HMM systems overpass CD-HMM/GMM approaches (CD-HMM/GMM lead to 13.0% and 31.7% WERs on the *dev* sets of Hausa and Wolof (initial) respectively[5]). These baseline performances show that our first Hausa ASR system can reach WER below 10%, even without any special modeling of the vowel length.

Table 1: *Results according to the baseline CD-DNN-HMM systems - Hausa and Wolof ASR - no modeling of vowel length.*

Language	**WER** (CER) **(%)**	
	dev	test
Hausa	8.0 (2.1)	11.3 (3.7)
Wolof (initial)	*27.2 (10.2)*	*33.6 (13.9)*
Wolof (cleaned)	20.5 (7.3)	24.9 (10.0)

The poor performance of the ASR system for Wolof can be explained by the high perplexity of the language model and by the higher OOV rates (0.19% of unknown words for Hausa and 4.2% for Wolof). Another problem with Wolof ASR is the lack of normalisation in the writing of words which penalizes both language model and WER (for this reason, we also display character error rate - CER). Nonetheless, as we can observe in table 1, when we decode only the cleaned data sets (see 3.1.3), we clearly improve the performance of the ASR system with an absolute gain of 6.7% on the *dev* set and 8.7% on the *test* set. The difference between *Wolof (intial)* and *Wolof (cleaned)* results shows that data cleaning is an additional and mandatory step when collecting read speech in languages with low literacy.

[4]The reason we consider the phoneme rather than the grapheme is because we want to analyse the phonetic realization in both languages. As we not only focus on the performance of the ASR systems but we also want to measure the duration of the vowels and its impact on the ASR system, a graphemic system was not considered in this study.

[5]Results not reported in the table.

4. Speed Perturbation

4.1. Corpus augmentation through speed perturbation

We used the following process to apply speed perturbation to our existing training corpora: first, we convert the initial files in *raw* format. Then, we slightly modify the initial sampling rate f to αf (with α=0.9 or α=1.1). Finally, we generate a wav file from the initial raw samples with new sampling rate αf.

Unlike VTLP, this simple processing produces a warped time signal which is faster or slower (depending on the value of α) than initial signal. Also, in the frequency domain, this leads to spectral contraction or dilatation (depending on α).

What is described above is probably very similar to the *speed* function of the Sox[6] audio manipulation tool (and used in [8]). However, using the function *speed* with a same factor α lead to different signal length compared to our algorithm above. In addition, we were not able to find a clear description of what the *speed* function does in Sox, so we decided to use both techniques in our experiments (later on, we will refer to α*-sampling* and to *sox/speed* to identify each approach). Anyway, while listening to the signals, the perceptive differences between α*-sampling* and *sox/speed* were small. Both approaches modify the pitch and spectral envelope of the signal simulating signals uttered by new speakers.

Both Hausa and Wolof training corpora were augmented using α*-sampling* (size x 3) or *sox/speed* (size x 3). In anticipation to ASR training, we decided to assign a new speaker label to all transformed signals corresponding to a same initial speaker. In other words, the number of speakers was multiplied by 3 when one transformation approach was used.

4.2. ASR results

We re-iterate the training pipeline with the augmented corpora and analyzed the impact on the ASR performance. Table 2 shows the performance obtained with the different data augmentation (speed perturbation) techniques on Hausa and Wolof.

Table 2: *Results according to the baseline CD-DNN-HMM systems - Hausa and Wolof ASR - with data augmentation (speed perturbation) - no modeling of vowel length so far.*

Language	**WER** (CER) **(%)**			
	α-sampling		***sox/speed***	
	dev	test	dev	test
Hausa	8.7 (2.4)	12.3 (3.9)	8.5 (2.3)	12.6 (3.9)
Wolof (cleaned)	20.3 (7.2)	24.4 (9.8)	20.0 (7.1)	24.3 (9.9)

As we can see, the difference between α*-sampling* and *sox/speed* is small. On the Wolof, the *sox/speed* method provides a slight improvement in comparison to the α*-sampling* method. Compared to the baseline presented in table 1 for the Wolof, we got an absolute gain of 0.5% on the *dev* set and 0.6% on the *test* set with the *sox/speed* method. On the contrary, neither approaches improve the performance of the ASR system for Hausa. One explanation might be that the speaker density on Hausa is already high (12.6 speakers per hour of train signal) while speaker density is much lower for Wolof (0.8 speakers per hour of train signal). The best absolute gain observed on Wolof (-0.6%) is similar to what was observed by [8] on English tasks: Librispeech (-0.4%), Ted-Lium (-0.5%) while bigger gains were observed on Switchboard (-1.4%). An analysis of the speaker density in these databases shows the same trend: data augmentation gains are bigger on Switchboard (density < 1 spk/h) than on Librispeech and Ted-Lium (density from 2 to 5 spk/h). However, this explanation should be considered with caution and further investigations are needed to confirm it.

[6]http://sox.sourceforge.net/sox.html

5. Vowel duration modeling in ASR

5.1. Empirical analysis of vowel length contrast

As said in Section 2.2, we wanted to empirically verify vowel length contrast before deciding to separate a given phone in two short/long units for ASR. For this, we trained new Wolof and Hausa acoustic models by forcing the system to represent this contrast (i.e. having different units in the phone set). While in Wolof the length mark is easy to annotate because it is marked on the orthography by a duplication of the grapheme (phones were marked with either a "_short" or a "_long" label), this is not the case in Hausa. In Hausa, the vowel length depends on several factors : the position of the vowel within the word (initial, middle, final), its position within the syllable, and also if the vowel is in pre-pausal position [23]. As the last two apply to very specific aspects of the language and even intrinsic values to the word, we have, for the moment, focused on the syllabic context only (phones marked as "_closed" or "_open" label). Based on these facts, we added length marks in the pronunciation dictionary according to the language treated. As a first step, we distinguished the length contrast for all the vowels of both languages. However, an analysis of the forced-alignments obtained showed that this contrast was not marked for all vowels. For instance, figure 1 shows that, for Wolof, the contrast is empirically observed for phoneme /o/ while it is not clearly marked for phoneme /i/ (which has few long occurences). This *machine assisted* analysis of vowel length contrast shows that its concrete realization depends on the vowel considered. For Wolof, the contrast was only observed for the following phonemes: /a/, /e/, /ɛ/, /o/ and /ɔ/[7]. For Hausa, the contrast was only clearly observed for two phonemes: /e/ and /o/ (see figure 2 which shows the distribution of the vowel lengths for /e/ and /o/ depending on their syllabic context). Moreover, for the three other Hausa vowels, the syllabic context (i.e.: closed *versus* open) alone did not allow a clear decision on their length and due to this constraint, only /e/ and /o/ were kept contrasted.

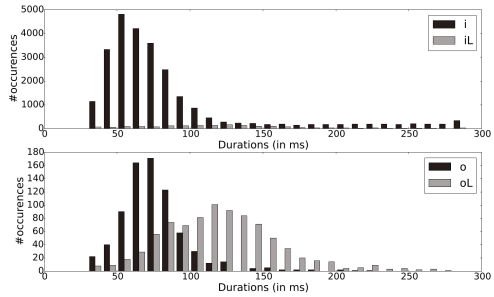

Figure 1: *Large scale empirical analysis of vowel length contrast for /i/ (no contrast observed) and /o/ (contrast observed) obtained by forced alignment - Wolof language.*

5.2. ASR experiments

Based on observations made in previous section, we trained new ASR systems for Hausa and Wolof (on the initial, not augmented corpus) taking into account vowel length contrast for a subset of vowels[8]. The length contrast labelling set used for each language is summarized in table 3[9].

[7] Due to space constraints, the analysis of all phonemes are not provided - figure 1 illustrates what was obtained for two vowels with or w/o length contrast.

[8] We trained an ASR system taking into consideration the length contrast for the whole vowel set of both languages and we got a worse score: 8.3% on the *dev* set for Hausa and 21.0% on the *dev* cleaned set for Wolof.

[9] For Hausa, we also have a few vowels labeled /e_unk/ and /o_unk/ since we still have some /e/ and /o/ non labeled, due to the large variability of the vowels phonetic realization depending on their position within the word.

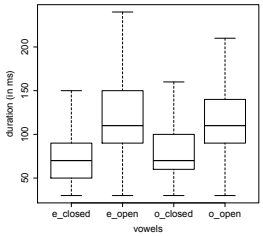

Figure 2: *Large scale empirical analysis of vowel length contrast for /e/ and /o/ (contrast observed) depending on their syllabic context - Hausa language.*

Table 3: *Summary of the vowel contrasts considered for each new ASR system trained.*

New system	Hausa ASR	Wolof ASR
# contrasts	2 ([e], [o])	5 ([a], [e], [ɛ], [o], [ɔ])
Label	_closed / _open / _unk	_short / _long

At this step, the pronunciation dictionary of Hausa is composed of 37 phonemes (instead of 33) and the one for Wolof is composed of 39 phonemes (instead of 34).
Using these new pronunciation dictionaries, we trained for each language, a new system with the same protocol as in 3.2. The triphone model for Hausa is trained by using 2,969 context-dependent states, the one for Wolof by employing 3,406 context-dependent states and 40K Gaussians. The performance of these new systems is shown in table 4. The first results in the table remind the WER/CER score for the baseline ASR systems (no vowel length modeling). Below this row, we present the score obtained when we modeled the length contrast existing in a vowel. Finally, the last row displays the scores when we combine both systems.

Table 4: *Results according to the new CD-DNN-HMM acoustic models - taking into account vowel length contrast.*

ASR systems	WER (CER) (%)			
	Hausa		**Wolof (cleaned)**	
	dev	test	dev	test
No vowel length modeling	8.0 (2.1)	11.3 (3.7)	20.5 (7.3)	24.9 (10.0)
Vowel length modeling	7.9 (2.1)	10.6 (3.5)	20.0 (7.0)	24.5 (9.8)
Combination of systems with/w-o length modeling	**7.8** (2.1)	**10.3** (3.6)	**19.1** (7.2)	**22.9** (9.7)

The consideration of the length contrasts on vowels, and specifically the selection of vowels on which the length contrast modeling was empirically observed, improves the performance of the ASR systems. Furthermore, the modeling of vowel length contrast seems to bring complementary information to the baseline since the system combination leads to significant ($p<0.01$) improvements on Hausa *test* and Wolof *dev* and *test*.

6. Conclusion

This paper proposed to model vowel length contrast to optimize ASR for two Sub-Saharan African languages: Hausa and Wolof. As a by-product, the acoustic models trained can be used for large scale phonetic analysis. While vowel length modeling was proven useful for both languages, data augmentation through speed perturbation only worked for the Wolof language. As a last experiment in Wolof, we combined both CD-DNN-HMM-based acoustic models generated from systems handling vowel duration and speed perturbation (*sox/speed*). We reached 18.9% of WER (CER=7.2%) for the cleaned *dev* set and 22.7% of WER (CER=9.5%) for the cleaned *test* set which is the best performance obtained so far on our Wolof LVCSR system.

7. Acknowledgements

This work was realized in the framework of the French ANR project ALFFA (ANR-13-BS02-0009).

8. References

[1] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Communication*, vol. 56, pp. 85–100, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2013.07.008

[2] J. K. Tamgno, E. Barnard, C. Lishou, and M. Richomme, "Wolof speech recognition model of digits and limited-vocabulary based on hmm and toolkit," in *Computer Modelling and Simulation (UKSim), 2012 UKSim 14th International Conference on*, March 2012, pp. 389–395.

[3] P. Newman, "The hausa language," *An Encyclopedic Reference Grammar'(New Haven & London)*, 2000.

[4] M. Paulin, J. Revaud, Z. Harchaoui, F. Perronnin, and C. Schmid, "Transformation pursuit for image classification," 2014.

[5] A. Y. Hannun, C. Case, J. Casper, B. C. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," *CoRR*, vol. abs/1412.5567, 2014. [Online]. Available: http://arxiv.org/abs/1412.5567

[6] N. Jaitly and E. Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," in *International Conference on Machine Learning (ICML) Workshop on Deep Learning for Audio, Speech, and Language Processing, 2013, Atlanta, June 16-21, 2013*.

[7] A. Ragni, K. M. Knill, S. P. Rath, and M. J. F. Gales, "Data augmentation for low resource languages," in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, 2014, pp. 810–814. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2014/i14_0810.html

[8] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, 2015, pp. 3586–3589. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2015/i15_3586.html

[9] V. R. Gadde, "Modeling word duration for better speech recognition," in *Proceedings of NIST Speech Transcription Workshop*, 2000.

[10] D. Povey, "Phone duration modeling for lvcsr," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1. IEEE, 2004, pp. I–829.

[11] J. Pylkkönen and M. Kurimo, "Duration modeling techniques for continuous speech recognition." in *INTERSPEECH*, 2004.

[12] T. Alumäe and R. Nemoto, "Phone duration modeling using clustering of rich contexts." in *INTERSPEECH*. Citeseer, 2013, pp. 1801–1805.

[13] T. Schlippe, E. G. K. Djomgang, N. T. Vu, S. Ochs, and T. Schultz, "Hausa large vocabulary continuous speech recognition." in *SLTU*, 2012, pp. 11–14.

[14] S. Nouguier Voisin, "Relations entre fonctions syntaxiques et fonctions sémantiques en wolof," Ph.D. dissertation, Lyon 2, 2002.

[15] J. L. Diouf, *Dictionnaire wolof-français et français-wolof*. KARTHALA Editions, 2003.

[16] G. Attardi and A. Fuschetto, "Wikipedia extractor," *Medialab, University of Pisa*, 2013.

[17] E. Gauthier, L. Besacier, S. Voisin, M. Melese, and U. P. Elingui, "Collecting resources in sub-saharan african languages for automatic speech recognition: a case study of wolof," *LREC*, 2016.

[18] D. Blachon, E. Gauthier, L. Besacier, G.-N. Kouarata, M. Adda-Decker, and A. Rialland, "Parallel speech collection for under-resourced language studies using the lig-aikuma mobile device app," *SLTU*, 2016.

[19] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," 2011.

[20] B. Kingsbury, "Lattice-based optimization of sequence classification criteria for neural network acoustic modeling," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2009, pp. 3761–3764.

[21] G. E. Hinton, "A practical guide to training restricted boltzmann machines," Dept. Computer Science, University of Toronto, UTML TR 2010-003, 2010.

[22] A. Fal, R. Santos, and J. L. Doneux, *Dictionnaire wolof-français: suivi d'un index français-wolof*. Karthala, 1990.

[23] R. M. Newman and V. J. v. Heuven, "An acoustic and phonological study of pre-pausal vowel length in hausa," *Journal of African Languages and Linguistics*, vol. 3, no. 1, pp. 1–18, 1981.