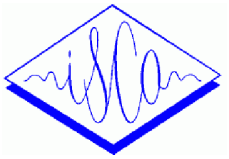

Third Workshop on Spoken Language
Technologies for Under-resourced Languages
Cape Town, South Africa
May 7-9, 2012

HAUSA LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION

Tim Schlippe, Edy Guevara Komgang Djomgang, Ngoc Thang Vu, Sebastian Ochs, Tanja Schultz

Cognitive Systems Lab, Karlsruhe Institute of Technology (KIT), Germany

ABSTRACT

We report on our efforts toward an LVCSR system for the African language Hausa. We describe the Hausa text and speech database recently collected as a part of our *GlobalPhone* corpus [1]. The data was complemented by a large collection of text data crawled from various Hausa websites. We achieve significant improvement by automatically substituting inconsistent or flawed pronunciation dictionary entries, including tone and vowel length information, applying state-of-the art techniques for acoustic modeling, and crawling large quantities of text material from the Internet for language modeling. A system combination of the best grapheme- and phoneme-based 2-pass systems achieves a word error rate of 13.16% on the development set and 16.26% on the test set on read newspaper speech.

Index Terms— speech recognition, rapid language adaptation, Hausa, African language

1. INTRODUCTION

Speech technology potentially allows everyone to participate in today's information revolution. Moreover, it can bridge language barrier gaps and facilitates worldwide business activities, simplifies life in multilingual communities, and alleviates humanitarian missions. To create speech processing systems, transcribed speech resources, large amounts of text for language modeling, and pronunciation dictionaries are of great importance. However, many languages still come with little or no speech and text resources. Africa itself has more than 2,000 languages [2] plus many different accents, e.g. there are more than 280 languages in Cameroon [3]. Building these resources for each language from scratch is a cumbersome and time consuming task. For only a few of Africa's many languages, speech processing technology has been analyzed and developed so far. For example, some Arabic dialects in North Africa have been explored in several DARPA projects. The Meraca Institute and South African universities spend much effort in investigating speech technologies in the Southern parts of the continent. They have developed systems for multiple Bantu languages [4]. In East Africa, the Djibouti Center for Speech Research and Technobyte Speech Technologies in Kenya explore speech technology for Afar (the second language of Djibouti) and Kiswahili [5]. To the best of our knowledge, in West Africa only one organization, the African Languages Technology Initiative (ALT-i) in Nigeria, has been investigating speech technology for Yoruba and Igbo [5]. We have collected Hausa speech and text data in Cameroon and developed an automatic speech recognition (ASR) system. Hausa is spoken in many countries located in West Africa and serves as a lingua franca there. Our Rapid Language Adaptation Toolkit (RLAT) [6] aims to significantly reduce the amount of time and effort involved in building speech processing systems for new languages and domains. It is envisioned to be achieved by providing tools that enable users to develop speech processing models, collect appropriate speech and text data to build these models, as well as evaluate the results allowing for iterative improvements. The purpose of this study is to apply RLAT to Hausa for collecting a large speech and text corpus in Cameroon, increase our knowledge of Hausa ASR, and further advance the language-dependent modules in RLAT and the ASR system to include the peculiarities of Hausa.

2. THE HAUSA LANGUAGE

Hausa is a member of the Chadic language family, which places it with the Semitic and Cushitic languages in the Afroasiatic language stock. With over 25 million speakers, it is widely spoken in West Africa [7]. The Hausa people are concentrated mainly in Northwestern Nigeria and in Southern Niger. The cities of this region - Kano, Sokoto, Zari, and Katsina, to name only a few, are among the largest commercial centers of sub-Saharan Africa. Hausa people also live in other countries of West Africa like Cameroon, Togo, Chad, Benin, Burkina Faso, and Ghana [8]. About one-fourth of Hausa words come from Arabic. We have recorded Hausa speech data in Cameroon. The spoken Hausa there is also influenced by French. Hausa's modern official orthography is a Latin-based alphabet called *boko*, which was imposed in the 1930s by the British colonial administration. It consists of 22 characters of the English alphabet (A/a, B/b, C/c, D/d, E/e, F/f, G/g, H/h, I/i, J/j, K/k, L/l, M/m, N/n, O/o, R/r, S/s, T/t, U/u, W/w, Y/y, Z/z) plus Ɓ/ɓ, Ɗ/ɗ, Ƙ/ƙ, 'Y/'y, and '. In many online newspapers, the characters Ɓ/ɓ, Ɗ/ɗ, and Ƙ/ƙ are mapped to B/b, D/d, and K/k. There are three lexical tones in Hausa, i.e. each of the five vowels /a/, /e/, /i/, /o/, and /u/ may have low tone, high tone, or falling tone [9].

Additionally, it is distinguished between short and long vowels which can also affect word meaning. Neither the vowel lengths nor the tones are marked in standard written Hausa. Hausa has also been written in *ajami*, a variant of the Arabic script, since the early 17th century. There is no standard system of using *ajami*. Therefore we have collected Hausa text data written in *boko*.

3. HAUSA RESOURCES

3.1. Text Corpus

To build a large corpus of Hausa text, we used RLAT [6] to crawl text from five websites as listed in Tab. 1, covering main Hausa newspaper sources in *boko*. RLAT enables the user to crawl text from a given webpage with different link depths. The websites were crawled with a link depth of 5 or 10, i.e. we captured the content of the given webpage, then followed all links of that page to crawl the content of the successor pages (link depth 2) and so forth until we reached the specified link depth. After collecting the Hausa text content of all pages, the text was cleaned and normalized in the following four steps: (1) Remove all HTML tags and codes, (2) remove special characters and empty lines, (3) identify and remove pages and lines from other languages than Hausa based on large lists of frequent Hausa words, and (4) delete duplicate lines. The websites were used to extract text for the language model (LM) and to select prompts for recording speech data for the training, development (*dev*), and evaluation (*test*) set.

Source	Websites
1	http://hausa.cri.cn
2	http://www.bbc.co.uk/hausa
3	http://www.dw-world.de/hausa
4	http://www.hausa.rfi.fr
5	http://www.voanews.com/hausa/news

Table 1. List of crawled Hausa Websites.

3.2. Speech Corpus

To develop and evaluate our Hausa recognizer, we collected Hausa speech data in *GlobalPhone* style[1] [1], i.e. we asked native speakers of Hausa in Cameroon to read prompted sentences of newspaper articles. As our web-based recording tool in RLAT turned out to be difficult to use as many sites in Cameroon did not provide Internet connection, we used the offline version. In total, the corpus contains 7,895 utterances spoken by 33 male and 69 female speakers in the age range of 16 to 60 years. All speech data was recorded with a headset microphone in clean environmental conditions. The data is sampled at 16 kHz with a resolution of 16 bits and stored in PCM encoding. The Hausa portion of the *GlobalPhone* database is listed in Tab. 2. Our speech data contains a variety of accents: Maroua, Douala, Yaoundé, Bafoussam, Ngaoundéré, and Nigeria. The dev set was used to determine the optimal parameters for our ASR system.

Set	Male	Female	#utterances	#tokens	Duration
Training	24	58	5,863	40k	6 h 36 mins
Development	4	6	1,021	6k	1 h 02 mins
Evaluation	5	5	1,011	6k	1 h 06 mins
Total	33	69	7,895	52k	8 h 44 mins

Table 2. Hausa *GlobalPhone* Speech Corpus.

4. BASELINE SPEECH RECOGNITION SYSTEM

Based on the International Phonetic Alphabet [9], we defined 33 Hausa phonemes as acoustic model units. The phone set consists of 26 consonants, 5 vowels, and 2 diphthongs. The 6.6 hours of the training set were used to train the acoustic models (AMs) of the Hausa speech recognizer. To rapidly build a baseline recognizer for Hausa, we applied RLAT [6] using a multilingual phone inventory for bootstrapping the system. This phone inventory *MM7* was trained from seven randomly selected *GlobalPhone* languages (Chinese, Croatian, German, English, Spanish, Japanese, and Turkish) [10]. To bootstrap the system, the Hausa phoneme models were initialized from the closest matches of the *MM7* inventory derived from an IPA-based phone mapping. We adopted the *GlobalPhone*-style preprocessing and used the selected *MM7* models as seed models to produce initial state alignments for the Hausa speech data. The preprocessing consists of feature extraction applying a Hamming window of 16ms length with a window overlap of 10ms. Each feature vector has 143 dimensions by stacking 11 adjacent frames of 13 coefficient Melscale Frequency Ceptral Coefficients (MFCC) frames. A Linear Discriminant Analysis (LDA) transformation is computed to reduce the feature vector size to 42 dimensions. The AM uses a fully-continuous 3-state left-to-right HMM. The emission probabilities are modeled by Gaussian Mixtures with diagonal covariances. For our context-dependent AMs with different context sizes, we stopped the decision tree splitting process at 500 triphones. After context clustering, a merge-and-split training was applied, which selects the number of Gaussians according to the amount of data. For all models, we use one global semi-tied covariance (STC) matrix after LDA. To model the tones, we apply the "Data-driven tone modeling" which had been successfully applied to the tonal language Vietnamese as described in [11]. In this method all tonal variants of a phoneme share one base model. However, the information about the tone is added to the dictionary in form of a tone tag. Our speech recognition toolkit allows to use these tags as questions to be asked in the context decision tree when building context-dependent

[1] GlobalPhone is a multilingual speech and text data collection in 20 languages available from ELRA (http://catalog.elra.info)

AMs. This way, the data decide during model clustering if two tones have a similar impact on the basic phoneme. If so, the two tonal variants of that basic phoneme would share one common model. In case the tone is distinctive (of that phoneme and/or its context), the question about the tone may result in a decision tree split, such that different tonal variants of the same basic phonemes would end up being represented by different models. For the vowel lengths, we apply the same technique. With the training transcriptions, we built a statistical 3-gram LM (*TrainTRL*) which contains their whole vocabulary (4k) plus 2k frequency-based selected words (see Tab. 5). It has a perplexity (PPL) of 282 and an out-of-vocabulary (OOV) rate of 4.7% on the dev set. The pronunciations for the 6k words were created in a rule-based fashion and were manually revised and cross-checked by native speakers. The performance of the baseline system is 23.49% on the dev set.

5. SYSTEM OPTIMIZATION

5.1. Pronunciation Dictionary Improvement

Since it is a cumbersome and error-prone task to create a pronunciation dictionary, we continuously improve our methods to automatically detect and substitute inconsistent or flawed entries. Furthermore, we analyzed the importance of tone and vowel length modeling for Hausa ASR.

5.1.1. Automatic rejection of inconsistent or flawed entries

We investigated different methods to filter erroneous word-pronunciation pairs and substitute the filtered pronunciations with more reliable ones:

1. Length Filtering (Len)
   (a) Remove a pronunciation if the number of grapheme and phoneme tokens differs more than a certain threshold.
2. Alignment Filtering (Eps)
   (a) Perform a grapheme-to-phoneme (g2p) alignment [12][13]. The alignment process involves the insertion of graphemic and phonemic nulls (epsilon) into the lexical entries of words.
   (b) Remove a pronunciation if the number of graphemic and phonemic nulls exceeds a threshold.
3. g2p Filtering after Length/Alignment Filtering ($G2P_{Len}$/$G2P_{Eps}$)
   (a) Train g2p models with "reliable" word-pronunciation pairs.
   (b) Apply the g2p models to convert a grapheme string into a most likely phoneme string.
   (c) Remove a pronunciation if the edit distance between the synthesized phoneme string and the pronunciation in question exceeds a threshold.

Dictionary	WER (%) on dev
Baseline (with tones and length)	23.49
Length Filtering (Len)	23.20
Alignment Filtering (Eps)	23.30
g2p Filtering after Length Filtering ($G2P_{Len}$)	**22.88**
g2p Filtering after Alignment Filtering ($G2P_{Eps}$)	23.15
Grapheme-based	22.52

Table 3. Automatic rejection of inconsistent or flawed entries.

The threshold for each filtering method depends on the mean (μ) and the standard deviation (σ) of the measure in focus (computed on all word-pronunciation pairs), i.e. the ratio between the numbers of grapheme and phoneme tokens in Len, the ratio between the numbers of graphemic and phonemic nulls in Eps, and the edit distance between the synthesized phoneme string and the pronunciation in question in $G2P_{Len}$ and $G2P_{Eps}$. Those word-pronunciation pairs whose resulting number is shorter than $\mu - \sigma$ or longer than $\mu + \sigma$ are rejected. We use the numbers of remaining word-pronunciation pairs to build new g2p models and apply them to the words with rejected pronunciations. Each filtering method substituted approximately 16% of the pronunciations. Tab. 3 shows that we are able to reduce the word error rate (WER) with all filtered pronunciation dictionaries compared to the baseline. $G2P_{Len}$ performed best and was selected for the tone and vowel length experiments. Additionally, we built a grapheme-based system which slightly outperforms all filtering methods.

Dictionary	WER (%) on dev
Phoneme-based (no tones, no vowel length)	24.33
Phoneme-based (no tones, vowel length)	23.15
Phoneme-based (tones, no vowel length)	23.06
Phoneme-based (tones, vowel length)	**22.88**

Table 4. Results with Tones and Vowel Lengths.

5.1.2. Tones and Vowel Lengths

We analyzed the importance of tone and vowel length modeling by including and excluding tone and vowel length information in the pronunciation dictionary. Tab. 4 indicates that best performance can be obtained modeling both (*Phoneme-based (tones, vowel length)*).

5.2. Language Model Improvement

We observed that the Hausa text provided on most websites is very limited. To improve the n-gram estimation in the LM and reduce the OOV rate, we crawled additional text corpora from one online newspaper (*http://hausa.cri.cn*) with longer crawling periods. After our text normalization steps, text with approximately 8 million tokens remained. By interpolating the invididual models built from the training transcriptions and

Language Model	dev / test	PPL	OOV	WER
TrainTRL (6k)	dev	281.7	4.68	22.88
	test	283.3	4.88	26.98
TrainTRL+Web (42k)	dev	154.7	0.51	**14.40**
	test	157.0	0.46	**17.83**

Table 5. LM Improvement (Additional Web Data).

the online newspaper, we created a new LM. The interpolation weights were tuned on the dev set transcriptions by minimizing the PPL of the model. We increased the vocabulary of the LM by selecting frequent words from the additional text material which are not in the transcriptions. A 3-gram LM with a total of 42k words (*TrainTRL+Web*) resulted in the lowest word error rates. Tab. 5 demonstrates that we were able to severely reduce the PPL, OOV rate, and WER using the additional web data.

5.3. Speaker Adaptation and System Combination

System combination methods are known to lower the WER of ASR systems [14]. They require the training of systems that are reasonably close in performance but at the same time produce an output that differs in their errors. This provides complementary information which leads to performance improvements. We trained speaker-independent (SI) and speaker-adaptive (SA) systems and obtained the necessary required variation with grapheme- and phoneme-based systems. Our experiments with a Confusion Network Combination (CNC) of the different systems resulted in a WER of 13.16% on the dev set. Fig. 1 gives an overview of our final system combination with the results of each system. On the test set we obtained a WER of 16.26%.

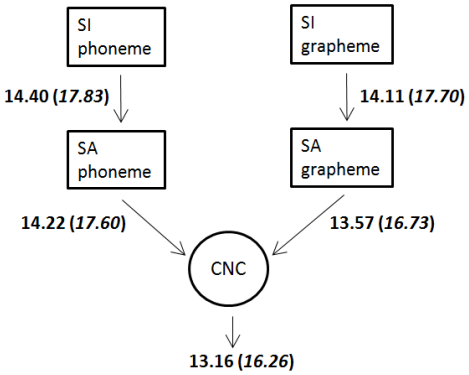

Fig. 1. System Combination Results (%) on dev (*test*) set.

6. CONCLUSION AND FUTURE WORK

We have described the development of a Hausa speech recognition system for large vocabulary. Hausa is the lingua franca in West Africa spoken by over 25 million speakers. We collected almost 9 hours of speech from 102 Hausa speakers reading newspaper articles. For language modeling, we collected a text corpus of roughly 8M words. After a rapid bootstrapping, based on a multilingual phone inventory, using RLAT, we improved the performance by carefully investigating the peculiarities of Hausa. The modeling of tones and vowel lengths performs better than omitting tone or vowel length information. We were able to improve the pronunciation dictionary quality with methods to filter erroneous word-pronunciation pairs. The initial recognition performance of 23.49% WER was improved to 13.16% on the dev set and 16.26% on the test set. Future work may concentrate on improving our pronunciation filtering methods and enhancing the LM with online newspapers in *ajami*.

7. REFERENCES

[1] T. Schultz, "GlobalPhone: A Multilingual Speech and Text Database Developed at Karlsruhe University," in *ICSLP*, 2002.

[2] B. Heine and D. Nurse, *African Languages: An Introduction*, Cambridge University Press, 2000.

[3] "Ethnologue," `http://www.ethnologue.com`.

[4] G. Pauw, G.-M-Schryver, L. Pretorius, and L. Levini, "Introduction to the Special Issue on African Language Technology," *Language Resources and Evaluation*, vol. 45, 2011.

[5] Tunde Adegbola, "Building Capacities in Human Language Technology for African Languages," in *AfLaT*, 2009.

[6] A. W. Black and T. Schultz, "Rapid Language Adaptation Tools and Technologies for Multilingual Speech Processing," in *ICASSP*, 2008.

[7] D. A. Burquest, "An Introduction to the Use of Aspect in Hausa Narrative," *Language in context: Essays for Robert E. Longacre, Shin Ja J. Hwang and William R. Merrifield (eds.)*, 1992.

[8] P. Koslow, *Hausaland: The Fortress Kingdoms*, The Kingdoms of Africa. Chelsea House Publishers, 1995.

[9] IPA, *Handbook of the International Phonetic Association: a guide to the use of the international phonetic alphabet*, Cambridge University Press, 1999.

[10] T. Schultz and A. Waibel, "Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition," *Speech Communication*, vol. 35, pp. 31–51, 2001.

[11] N. T. Vu and T. Schultz, "Vietnamese Large Vocabulary Continuous Speech Recognition," in *ASRU*, 2009.

[12] O. Martirosian and M. Davel, "Error Analysis of a Public Domain Pronunciation Dictionary," in *PRASA*, 2007, pp. 13–18.

[13] A. W. Black, K. Lenzo, and V. Pagel, "Issues in Building General Letter to Sound Rules," in *ESCA Workshop on Speech Synthesis*, 1998.

[14] S. Stüker, C. Fügen, S. Burger, and M. Wölfel, "Cross-System Adaptation and Combination for Continuous Speech Recognition: The Influence of Phoneme Set and Acoustic Front-End," in *Interspeech*, 2006.