

# Hausa Large Vocabulary Speech Recognition

Tim Schlippe

Edy Guevara Komgang Djomgang

Ngoc Thang Vu

Sebastian Ochs

Tanja Schultz

Cape Town, South Africa

07 May 2012

# Outline

1. Motivation
2. The Hausa Language
3. Hausa Resources
  - 3.1 Text Corpus
  - 3.2 Speech Corpus
4. Baseline Speech Recognition System
5. System Optimization
  - 5.1 Pronunciation Dictionary Improvement
    - 5.1.1 Automatic rejection of inconsistent or flawed entries
    - 5.1.2 Tones and vowel lengths
  - 5.2 Language Model Improvement
  - 5.3 Speaker Adaptation and System Combination
6. Conclusion

# 1. Motivation

- Speech technology ...
  - potentially allows everyone to participate in today's information revolution,
  - can bridge language barrier gaps,
  - facilitates worldwide business activities,
  - simplifies life in multilingual communities,
  - alleviates humanitarian missions.

# 1. Motivation

- Africa itself ...
  - has more than 2,000 languages (*Heine and Nurse, 2000*) (e.g. there are more than 280 languages in Cameroon ([www.ethnologue.com](http://www.ethnologue.com))).
  - plus many different accents
- For only a small fraction of Africa's many languages, speech technology has been analyzed and developed so far
- We have collected speech and text data in Cameroon for the West African language Hausa as a part of our *GlobalPhone* corpus (*Schultz, 2002*) and developed an automatic speech recognition system.

## 2. The Hausa Language

- Why Hausa?
  - Lingua franca in many countries
  - With over 25 million speakers, it is widely spoken in West Africa (*Burquest, 1992*)
  - Hausa speakers according to the Summer Institute of Linguistics (SIL):
    - 18.5 million in Nigeria (1991),
    - 5 million in Niger (1998),
    - 489k in Sudan (2001),
    - 23.5k in Cameroon (1982),
    - Benin, Burkina Faso, Ghana, Togo, Chad (*Koslow, 1995*)
  - Online text resources available
  - Phoneme set defined by International Phonetic Association (IPA) (*IPA, 1999*)

## 2. The Hausa Language

- Classification: Afro-Asiatic, Chadic, West, A, A.1
- Alphabet:
  - *ajami* (based on Arabic Alphabet), e.g. „**هوس**“
  - *boko* (based on Latin Alphabet), e.g. „Hausa“
    - 22 characters of the English Alphabet plus **B/b**, **D/d**, **K/k**, 'Y/y' or **Y/y**, and ' '
    - In online newspapers: **B/b**, **D/d**, **K/k** → **B/b**, **D/d**, **K/k**
- Pronunciation characteristics:
  - 3 lexical tones (low, high, falling) (*IPA, 1999*), e.g. wuya
    - wuyá → difficulty
    - wúya → neck
  - Vowel lengths (short, long), e.g. gari
    - garî → town
    - ga:rî → flour

# 3. Hausa Resources

## 3.1 Text Corpus

- Crawling text from 5 main online newspapers in *boko* using the Rapid Language Adaptation Toolkit (RLAT) (*Black and Schultz, 2008*)

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

- Text Normalization
  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
  4. Delete duplicate lines
- Select prompts to record speech data for the training, development, and evaluation set and extract text for the language model

# 3. Hausa Resources

## 3.2 Speech Corpus

- Speech data collection in *GlobalPhone* style (*Schultz, 2002*), i.e. we asked native speakers of Hausa to read prompted sentences of newspaper articles.
- Offline audio recorder
- 16 kHz sampling rate with 16 bit quantization
- Close talk microphone (noise cancellation microphone, NC-185VM)





# 3. Hausa Resources

## 3.2 Speech Corpus - Challenges

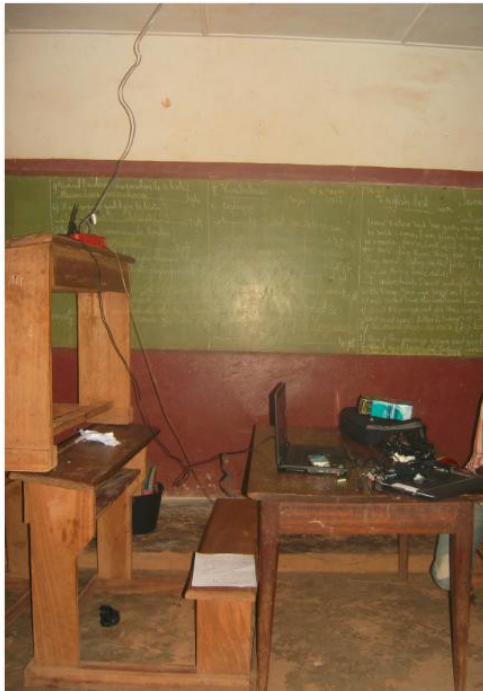
- Social factors
  - The majority of Hausa people is Muslim (95% of recorded speakers)
  - For Muslims close connection between work and religion
  - Most Muslim female speakers had to ask their husband or father for the permission to do the recording



# 3. Hausa Resources

## 3.2 Speech Corpus - Challenges

- Technical difficulties
  - Noisy environments:  
Big cities, restaurants, offices, at home, meeting halls
  - Bad infrastructure (electricity)





# 3. Hausa Resources

## 3.2 Speech Corpus

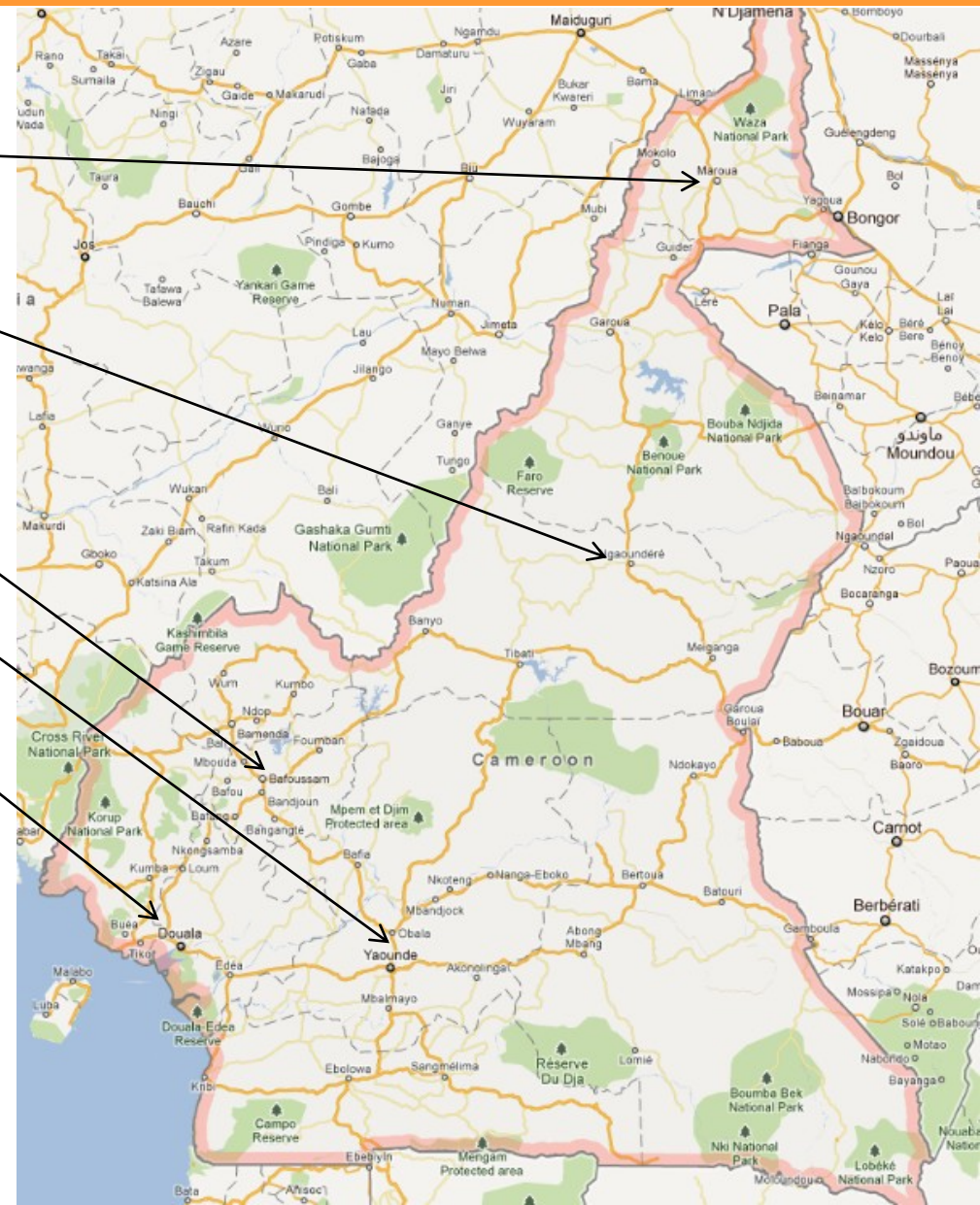
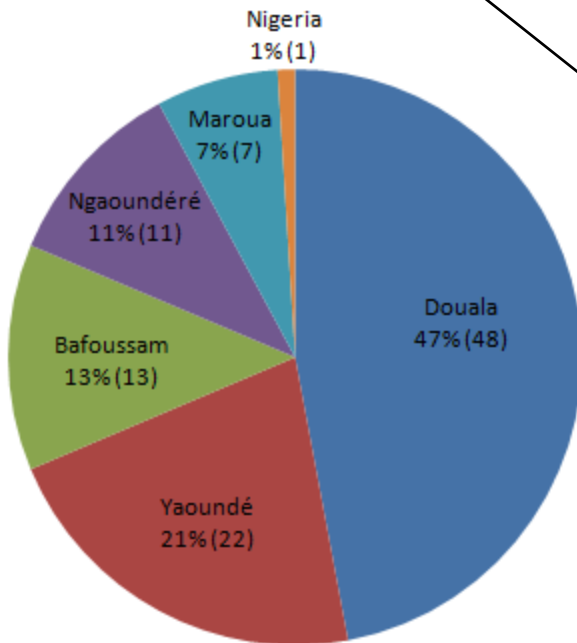
Maroua

Ngaoundéré

Bafoussam

Yaoundé

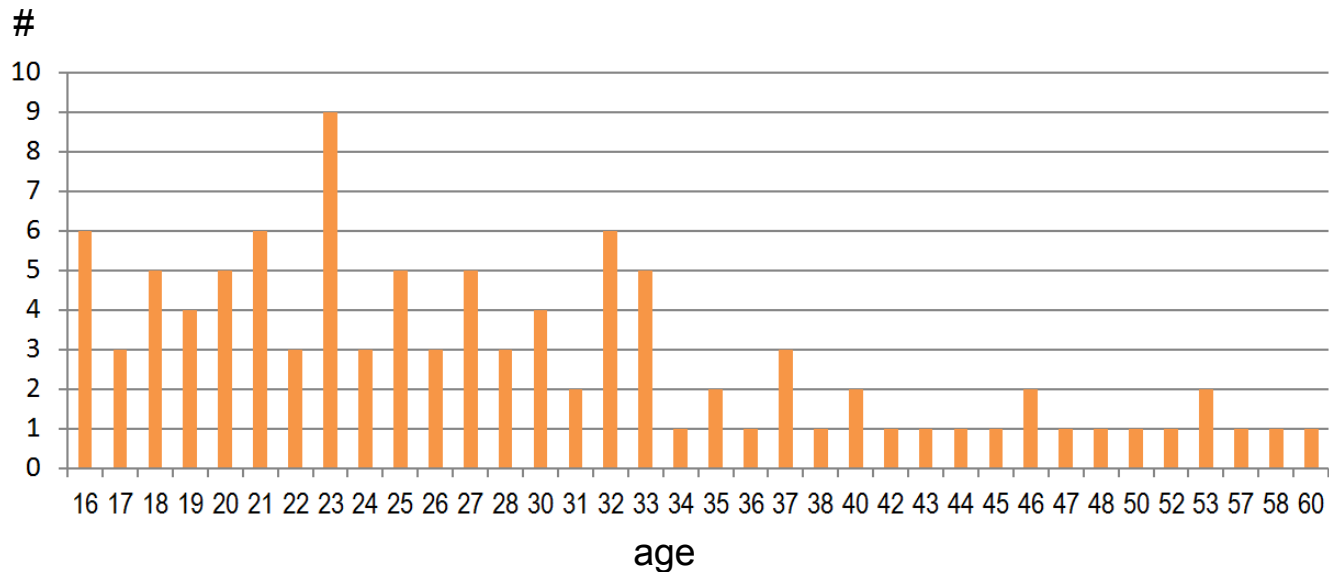
Douala



# 3. Hausa Resources

## 3.2 Speech Corpus

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



## 4. Baseline Speech Recognition System

- 33 Hausa phonemes (26 consonants, 5 vowels, 2 diphthongs)
- 6.6 hours to train acoustic models
- Bootstrapping with RLAT using multilingual phone inventory *MM7* (*Schultz and Waibel, 2001*)
  - *MM7* trained from 7 randomly selected *GlobalPhone* languages
  - Selected *MM7* models as seed models to produce initial state alignments for the Hausa speech data
- Preprocessing:
  - Feature extraction with Hamming window of 16 ms length, window overlap of 10 ms
  - Each feature vector has 143 dimensions (11 adjacent frames x 13 MFCC frames)
  - Linear Discriminant Analysis (LDA) → feature vector size: 42 dims.

# 4. Baseline Speech Recognition System

- Acoustic Model (AM):
  - Fully-continuous 3-state left-to-right HMM
  - Emission probabilities are modeled by Gaussian Mixtures with diagonal covariances
  - Context-dependent AM: decision tree splitting stopped at 500 triphones
  - For all AMs one global semi-tied covariance matrix after LDA
  - Data-driven tone modeling (DDTM) (*Vu and Schultz, 2009*)
    - Use a tone tag in pronunciation dictionary and add tag as question in clustering procedure
    - Data decide during model clustering if two tones have a similar impact on the basic phoneme
    - If so → share 1 common model;  
Otherwise → decision tree split → 2 different models
  - Same technique for the vowel lengths (Data-driven lengths modeling)

## 4. Baseline Speech Recognition System

- Language Model (LM):
    - 3-gram
    - Vocabulary size: 6k (4k training transcriptions + 2k frequent)
    - Perplexity (PPL): 282
    - Out-of-vocabulary (OOV) rate: 4.7%
  - Pronunciation Dictionary
    - Pronunciations created in a rule-based fashion, then manually revised and cross-checked by native speakers
      - Initial rules based on 200 word-pronunciation pairs from Peter Ladefoged (<http://archive.phonetics.ucla.edu/Language/HAU/hau.html>)
      - Then manual checks by different native speakers
- Performance of *baseline* on dev set: **23.49%**

# 5. System Optimization

## 5.1 Pronunciation Dictionary Improvement

### 5.1.1 Automatic rejection of inconsistent or flawed entries (1)

#### 1. Length Filtering (*Len*)

- a. Remove a pronunciation if the number of grapheme and phoneme tokens differs more than a threshold.

#### 2. Alignment Filtering (*Eps*) (according to (*Martirosian and Davel, 2007*))

- a. Perform a grapheme-to-phoneme (g2p) alignment (*Black et al., 1998*)
  - 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.

H		a		u		s		a	
H_h	(	_		H_aU		H_s		H_al	



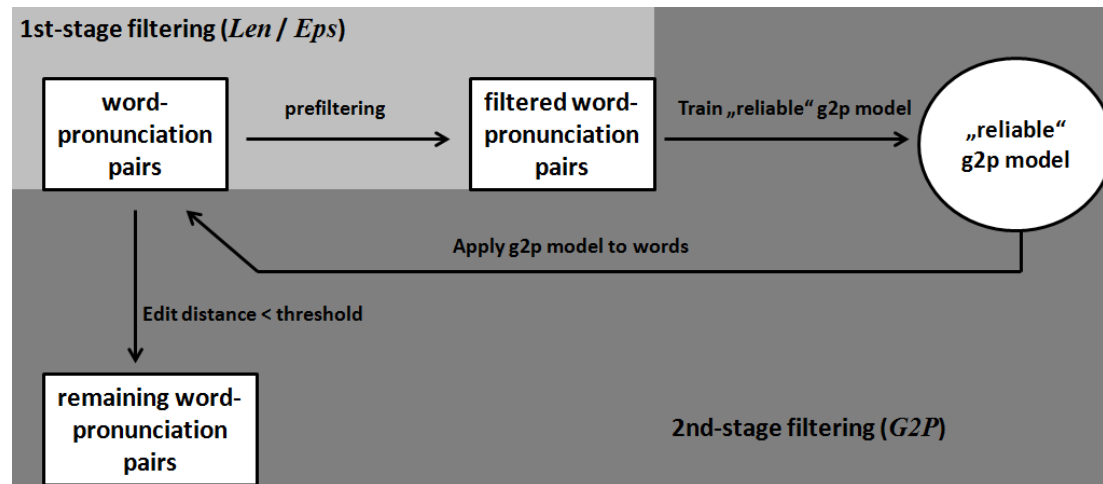
# 5. System Optimization

## 5.1 Pronunciation Dictionary Improvement

### 5.1.1 Automatic rejection of inconsistent or flawed entries (1)

#### 3. g2p Filtering after Length/Alignment Filtering ( $G2P_{Len}/G2P_{Eps}$ )

- Train g2p models with “reliable” word-pronunciation pairs.
- Apply the g2p models to convert a grapheme string into a most likely phoneme string.
- Remove a pronunciation if the edit distance betw. the synthesized phoneme string and the pronunciation in question exceeds a threshold.




# 5. System Optimization

## 5.1 Pronunciation Dictionary Improvement

### 5.1.1 Automatic rejection of inconsistent or flawed entries (2)

- The threshold for each filtering method depends on
  - the *mean* ( $\mu$ ) and
  - the *standard deviation* ( $\sigma$ ) of the measure in focus.
- Those word-pronunciation pairs whose resulting number is shorter than  $\mu - \sigma$  or longer than  $\mu + \sigma$  are rejected (~16% with each filtering methods).
- We built new grapheme-to-phoneme (g2p) models with the remaining word-pronunciation pairs and applied them to the words with rejected pronunciations



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

- 2.67% relative improvement with  $G2P_{Len}$  but still room for improvement (Grapheme-based: 22.52%)

# 5. System Optimization

## 5.1 Pronunciation Dictionary Improvement

### 5.1.2 Analysis of importance of tones and vowel length information

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)	<b>22.88</b>

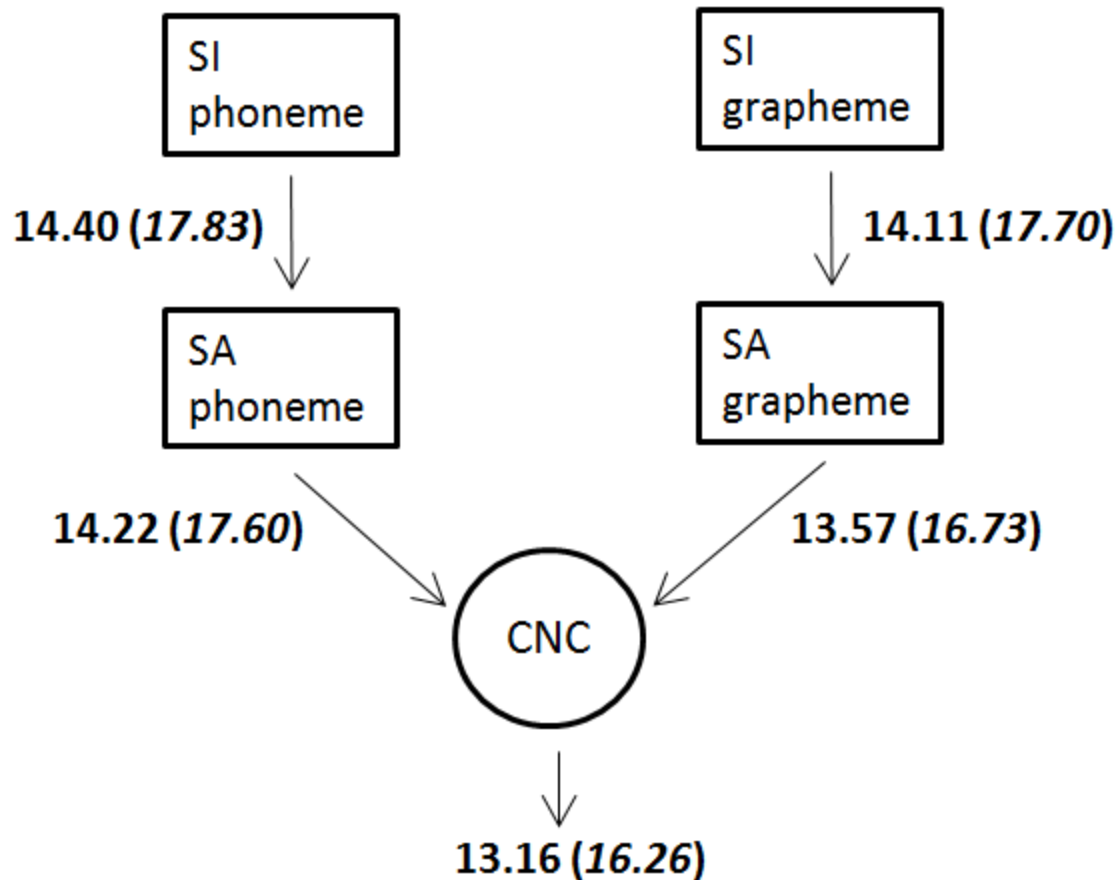
## 5.2 Language Model Improvement

- Crawling additional text corpora (8 million tokens from <http://hausa.cri.cn>)
- Increase vocabulary size with frequent words (42k resulted in best WER)

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	<b>14.40</b>
	test	157.0	0.46	<b>17.83</b>

# 5. System Optimization

## 5.3 Speaker Adaptation and System Combination



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

## 6. Conclusion

- Development of a Hausa speech recognition system for large vocabulary
- Hausa is a 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 automatic speech recognition, the modeling of tones and vowel lengths performs better than omitting tone or vowel length information
- We improved 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.

Thanks for your interest!

Baie dankie!

# 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.