Automatic Speech Recognition For African Languages With Vowel Length Contrast

Elodie Gauthier*, Laurent Besacier*, Sylvie Voisin**

*LIG, CNRS and Univ. Grenoble Alpes (France)
**DDL, CNRS and Univ. Lyon 2 (France)
Context: ALFFA project (1/2)

- African Languages and Information Technologies
- Address under-resourced languages from Africa
  - Focus on West Africa
    - Hausa, Wolof, Fulfulde, Zarma, Bambara
  - 2 East African languages
    - Amharic, Swahili
- Data collection methodology
- ASR (speech-to-text) and TTS (text-to-speech)
- French Partners: LIG (Grenoble), LIA (Avignon), DDL (Lyon), Voxygen (Lannion), http://alffa.imag.fr
This paper focuses on ASR

Four languages covered so far in ASR

- Hausa, Wolof, Amharic, Swahili

| Task                        | WER % |
|-----------------------------|-------|
| Swahili Broadcast News      | 20.7  |
| Hausa Read Speech           | 10.0  |
| Amharic Read Speech         | 8.7   |
| Wolof Read Speech (under dev.) | 28.6  |

Data and Kaldi [1] scripts released on Github

https://github.com/besacier/ALFFA_PUBLIC

[1] D. Povey & al., “The kaldi speech recognition toolkit,” IEEE ASRU 2011.
2 African languages largely spoken in the West of the continent

Vehicular languages also spoken as a 2d or 3d language

Both have length contrast at the phone level

- Contrast of vowels (Hausa) and contrast of vowels+consonants (Wolof)
- Contrast changes the meaning of a word
- Literature mention contrast for 5 vowels in Hausa and 7 vowels in Wolof

Length mark in text?

- Hausa: no (vowel length depends on its position into the syllable)
  - bos (bus) => /o/ expected short
  - aboki (friend) => /o/ expected long

- Wolof: yes (duplication of the graphemes)
  - kaar (bus) => /a/ expected long
  - xarit (friend) => /a/ expected short
Vowel length contrast in Hausa and Wolof (2/3)

Related works in ASR duration models

- [1] rescoring N-best lists with duration models
- [2] modeled duration at word and phone levels + lattice rescoring
  - WER reduction limited
- [3] proposed several durations models in Finish and [4] in Estonian
  - Moderate WER reduction

Comment: these approaches do not challenge the duration contrast phenomenon while for under-documented languages, it might be important to verify its empirical realization

[1] Gadde, V.R. Modeling word duration for better speech recognition. In: Proceedings of NIST Speech Transcription Workshop. 2000.
[2] Povey, D. Phone duration modeling for lvcsr. In: Proceedings IEEE ICASSP’04. 2004.
[3] Pylkkonen, J. & al. Duration modeling techniques for continuous speech recognition. In: INTERSPEECH. 2004.
[4] Alumäe, T. Phone duration modeling using clustering of rich contexts. In: INTERSPEECH. 2013.
Taking into account length contrast in Hausa and Wolof

- Train ASR systems with and w/o vowel length modeling
  - Different phone inventories and lexicons
- Combine these systems (instead of re-scoring framework)
- Vowel length modeling in Wolof
  - Contrastive phone units for short and long vowels (a gives /a_short/ and aa gives /a_long/)
- Vowel length modeling in Hausa
  - Contrastive phone units for phones if they are in a closed or open syllable (o gives /o_closed/ or /o_open/ depending on its position in the syllable)
  - Sometimes, not decidable: /o_unk/ label
Experiments & Results (1/4)

Data sets
- Hausa Globalphone Corpus [1]
- Wolof in-house Corpus [2]

Wolof
- 21h of read speech recorded in Dakar (Senegal), 18 speakers (10+8)
- In-house docs + text data crawled from the Web (but very few!!)

| Task                  | LM corpus (#words) | AM training (h) | Eval. Set (h) |
|-----------------------|--------------------|-----------------|---------------|
| Hausa read speech     | 8M                 | 7h              | 1h            |
| Wolof read speech     | 0.6M               | 17h             | 2h            |

[1] T. Schlippe & al. “Hausa large vocabulary continuous speech recognition.” In: SLTU, 2012.
[2] E. Gauthier & L. Besacier “Collecting resources in sub-saharan african languages for automatic speech recognition: a case study of wolof;” LREC, 2016.
Experiments & Results (1/4)

**ASR Systems**
- Kaldi CD-GMM-HMM and CD-DNN-HMM systems
- 2.9k CD states for Hausa, 3.4k CD states for Wolof
- Kaldi scripts made available
  - https://github.com/besacier/ALFFA_PUBLIC/tree/master/ASR/HAUSA
  - https://github.com/besacier/ALFFA_PUBLIC/tree/master/ASR/WOLOF

**Baseline results (no length modeling)**

| Task           | WER (GMM-HMM) | WER (DNN-HMM) |
|----------------|---------------|---------------|
| Hausa read speech | 13.0 %        | 8.0%          |
| Wolof read speech     | 31.7 %        | 27.2 %        |
Empirical verification of vowel length contrast in Hausa

- 5,863 training sentences forced-aligned
- Measured Hausa vowel length in closed vs open syllabic context
- Differences in vowel length only observable for /e/ and /o/

For /a/, /i/ and /u/ contrast was not empirically observed
Hausa Results

- Only /e/ and /o/ contrasted in the lexicon with _closed/_open/_unk labels.
- Slight gain, but not significant.
- Modeling contrast for all vowels degrades the results (not reported here).

| Method              | Hausa WER (DNN) |
|---------------------|-----------------|
| No length modeling  | 8.0 %           |
| Length modeling     | 7.9 %           |
| Combination         | 7.8 %           |
Wolof Results

- All 7 vowels contrasted in the lexicon with _short/_long
- Slight gain with combination only
- Did not verify the length contrast for Wolof at the time of paper submission (see next slide)

| Method                | Hausa WER (DNN) | Wolof WER (DNN) |
|-----------------------|-----------------|-----------------|
| No length modeling    | 8.0 %           | 27.2 %          |
| Length modeling       | 7.9 %           | 27.7 %          |
| Combination           | 7.8 %           | 26.3 %          |
**Post-submission experiments (1/1)**

Updated results on Wolof

- Only 5 vowels for which contrast was empirically verified
- Wolof eval. set cleaned since then (from 2,000 to 1,120 utterances)

| Method                | Hausa WER (DNN) | Wolof WER (DNN) |
|-----------------------|-----------------|-----------------|
| No length modeling    | 8.0 %           | 20.5 %          |
| Length modeling       | 7.9 %           | 20.0 %          |
| Combination           | 7.8 %           | 19.1 %          |
First LVCSR system developed for Wolof language

- Train ASR systems with and w/o vowel length modeling
- Complementarity of systems but small performance gains
- Acoustic models with length contrast useful for phonetic studies or TTS
  - Length contrast phenomenon not empirically observed on all vowels, even on read speech
  - What about spontaneous speech?

ALFFA project outcomes

- Kaldi ASR systems and resources online
- Started to work on Pulaar Fulfulde