Language Identification and Multilingual Speech Recognition using Discriminatively Trained Acoustic Models

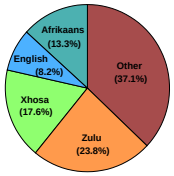
Thomas Niesler [†] and **Daniel Willett** [‡]

[†] Department of Electronic Engineering, University of Stellenbosch, South Africa

[‡] Harman/Becker Automotive Systems, Ulm, Germany

1. Background

- Language identification (LID) is key to spoken dialogue systems that must function in multilingual environments
- South Africa officially recognises 11 languages
- Most citizens are fluent in more than one language
- We focus on four languages that together account for 63% of mother−tongue speakers: Afrikaans South African English, Xhosa and Zulu



2. Databases

- Telephone speech annotated both orthographically and phonetically
- Mix of spontanous and read speech

Database name	Speech (hours)	No. of speakers	Phone types	Phone tokens
Afrikaans	6.18	234	84	180,904
English	6.02	271	73	167,987
Xhosa	6.98	219	107	177,843
Zulu	10.87	203	101	285,501

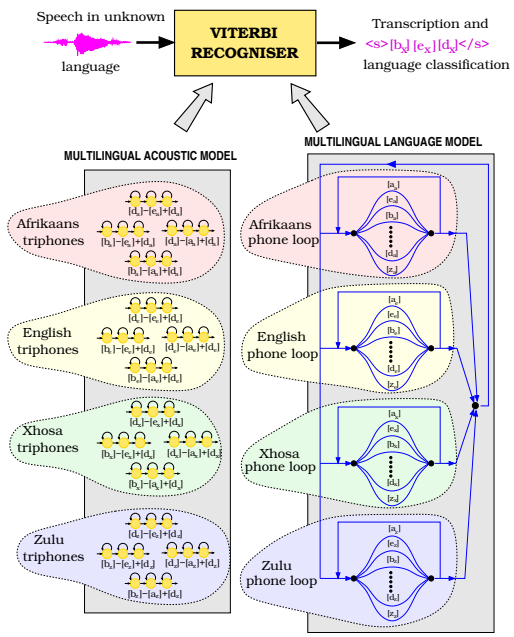
Training set

Database name	Development test		Evaluation test	
	Speech (mins)	No. of speakers	Speech (mins)	No. of speakers
Afrikaans	15.3	12	24.4	20
English	14.2	10	24.0	18
Xhosa	15.3	10	26.8	17
Zulu	16.8	10	27.1	16

Test sets

- Balance of male/female and mobile/fixed−line speakers
- Due to code−mixing, each database may contain words in other languages

3. System Architecture



- Speech parameterised as MFCCs, 1st and 2nd differentials, per−utterance CMN
- Diagonal covariance 8−mixture cross−word triphone models
- Baseline maximum−likelihood HMMs trained by embedded Baum−Welch reestimation

4. Discriminative Training

- Discriminitive training approaches aim to maximise:

$$\phi_{\text{disc}} = \underset{\phi \in \Phi}{\operatorname{argmax}} \prod_{u=1}^{U} \frac{p_\phi(X_u \mid L_u)}{p_\phi(X_u)}$$

$$p_\phi(X_u \mid L_u) \approx p_\phi(X_u \mid W_u)$$ Reference transcription (eq. 1)

$$p_\phi(X_u) \approx \max_W p_\phi(X_u \mid W)$$ Viterbi best, any language (eq. 2)

OR

$$p_\phi(X_u) \approx \max_{W,\, L_W \neq L_u} p_\phi(X_u \mid W)$$ Viterbi best, wrong language (eq. 3)

ϕ are HMM parameters
L_u is language of u_{th} utterance
X_u are acoustic data for u_{th} utterance
W_u orthographic transcription of u_{th} utterance

- Refine ML baseline by two iterations of discriminative training using Extended Baum−Welch algorithm
- HD1 shows good improvement on training set LID but not on test−set
- HD2 shows LID improvement on both training an test−sets
- ASR performance hardly affected

Model set	Configuration	LID accuracy (%)			Phone error rate (%)	
		Train	Dev	Eval	Dev	Eval
HML	ML baseline	87.13	81.49	79.69	46.42	45.85
HD1	DT eq. 1 & eq. 2	88.58	81.61	79.31	46.62	46.08
HD2	DT eq. 1 & eq. 3	87.22	82.68	79.92	46.17	45.97

5. Utterance length

- Average length is short: 2.1s for Afrikaans and English 2.6s for Xhosa and Zulu

Model set	Shorter utterances		Longer utterances	
	Dev	Eval	Dev	Eval
HML	79.42	79.36	84.51	80.19
HD1	79.73	78.79	84.51	80.09
HD2	81.29	79.55	84.98	80.28

- HD2 improves LID for both short and long utterances
- Discriminative training has narrowed gap between short and long utterances

5. Code mixing

- English often used to cite numbers, dates and times in modern Xhosa and Zulu
- 27.9% of Xhosa and 31.2% of Zulu test utterances contain 75% or more English words

Model set	Code−mixed utts		Remaining utts	
	Dev	Eval	Dev	Eval
HML	77.78	70.57	81.91	81.08
HD1	77.78	69.43	82.13	80.82
HD2	80.42	69.70	83.05	81.39

- LID is less accurate for code−mixed utterances
- Discriminative training worsens LID performance for code−mixed utterances

Presented at: ISCA Tutorial and Research Workshop (ITRW) on Multilingual Speech and Language Processing (Multiling−2006) , 9−11 April 2006 , Stellenbosch , South Africa