Speech Processing for Unwritten Languages

Alan W Black

Language Technologies Institute

Carnegie Mellon University

ISCSLP 2016 – Tianjin, China

Speech Processing for Unwritten Languages

Joint work with
Alok Parlikar, Sukhada Parkar,
Sunayana Sitaram, Yun-Nung (Vivian) Chen,
Gopala Anumanchipalli, Andrew Wilkinson,
Tianchen Zhao, Prasanna Muthukumar.
Language Technologies Institute
Carnegie Mellon University

ISCSLP 2016 – Tianjin, China

Speech Processing

- *The major technologies:*
  - *Speech-to-Text*
  - *Text-to-Speech*
- *Speech processing is **text** centric*

Overview

- Speech is not spoken text
- With no text what can we do?
  - Text-to-speech without the text
  - Speech-to-Speech translation without text
  - Dialog systems for unwritten languages
- Future speech processing models

Speech vs Text

- Most languages are not written
  - Literacy is often in another language
  - e.g. Mandarin, Spanish, MSA, Hindi
  - vs, Shanghaiese, Quechua, Iraqi, Gujarati
- Most writing systems aren't very appropriate
  - Latin for English
  - Kanji for Japanese
  - Arabic script for Persian

Writing Speech

- Writing is not for speech its for writing
- Writing speech requires (over) normalization
  - “gonna” → “going to”
  - “I'll” → “I will”
  - “John's late” → “John is late”
- Literacy is often in a different language
  - Most speakers of Tamil, Telugu, Kannada write more in English than native language
- Can try to force people to write **speech**
  - Will be noisy, wont be standardized

Force A Writing System

- Less well-written language processing
- Not so well defined
  - No existing resources (or ill-defined resources)
  - Spelling is not-well defined
- Phoneme set
  - Might not be dialect appropriate (or archaic)
  - (Wikipedia isn't always comprehensive)
- But what if you have (bad) writing and audio
  - Writing and Audio

Grapheme Based Synthesis

- Statistical Parametric Synthesis
  - More robust to error
  - Better sharing of data
  - Less instance errors
- From ARCTIC (one hour) databases (clustergen)
  - This is a pen
  - We went to the church and Christmas
  - Festival Introduction

Other Languages

- Raw graphemes (G)
- Graphemes with phonetic features (G+PF)
- Full knowledge (Full)

	G	G+PF	Full
English	5.23	5.11	4.79
German	4.72	4.30	4.15
Inupiaq	4.79	4.70	
Konkani	5.99	5.90	

Mel-cepstral Distortion (MCD) lower is better

Unitran: Unicode phone mapping

- Unitran (Sproat)
  - Mapping for all unicode characters to phoneme
  - (well almost all, we added Latin++)
  - Big table (and some context rules)
    - Grapheme to SAMPA phone(s)
  - (Doesn't include CJK)
  - Does cover all other major alphabets

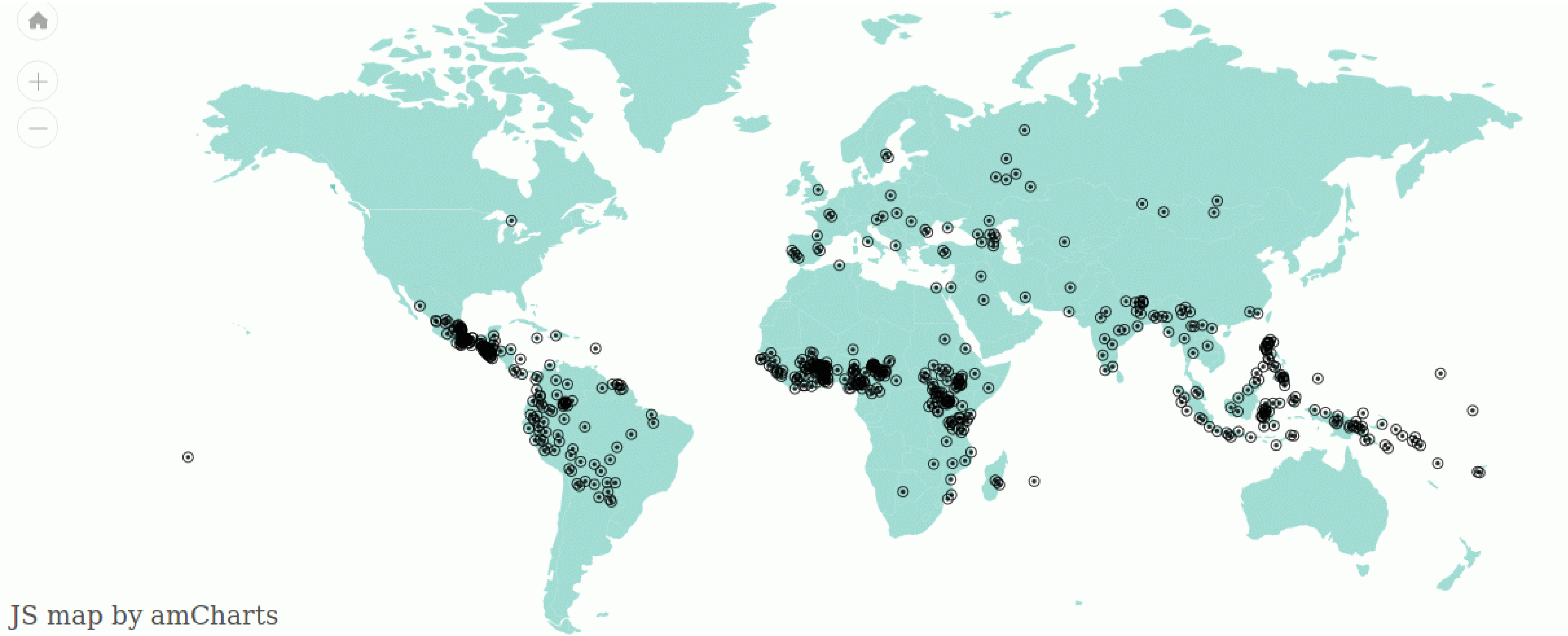
More Languages

- Raw graphemes
- Graphemes with phonetic features (Unitran)
- Full knowledge

	G	Unitran	Full
Hindi	5.10	5.05	4.94
Iraqi	4.77	4.72	4.62
Russian	5.13	4.78	
Tamil	5.10	5.04	4.90
Dari	4.78	4.72	

Wilderness Data Set

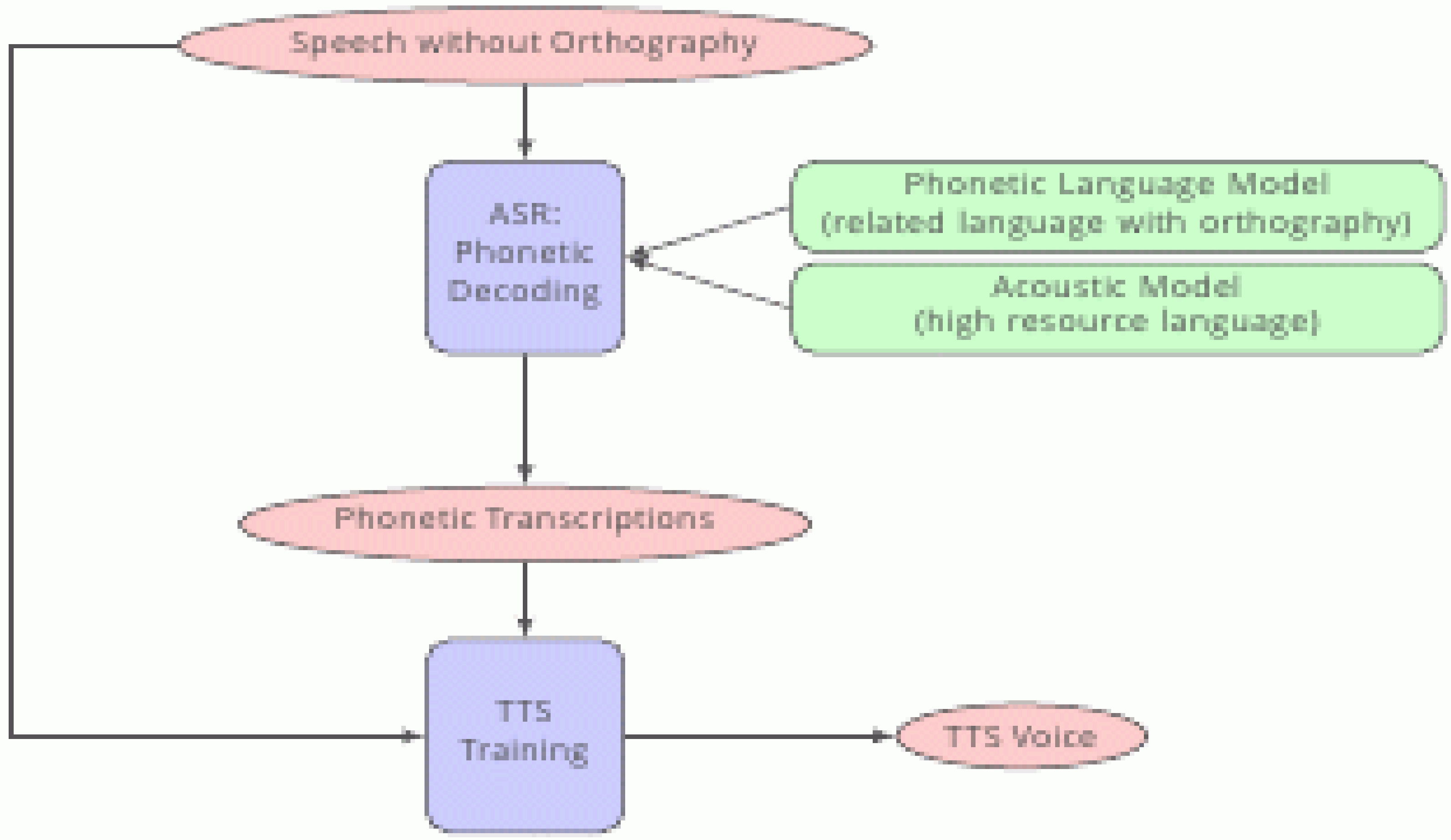
- 600+ Languages: 20 hours each
  - Audio, pronunciations, alignments
  - ASR and TTS
  - From Read Bibles.



TTS without Text

- Let's derive a writing system
  - Use cross-lingual phonetic decoding
  - Use appropriate phonetic language model
- Evaluate the derived writing with TTS
  - Build a synthesizer with the new writing
  - Test synthesis of strings in that writing

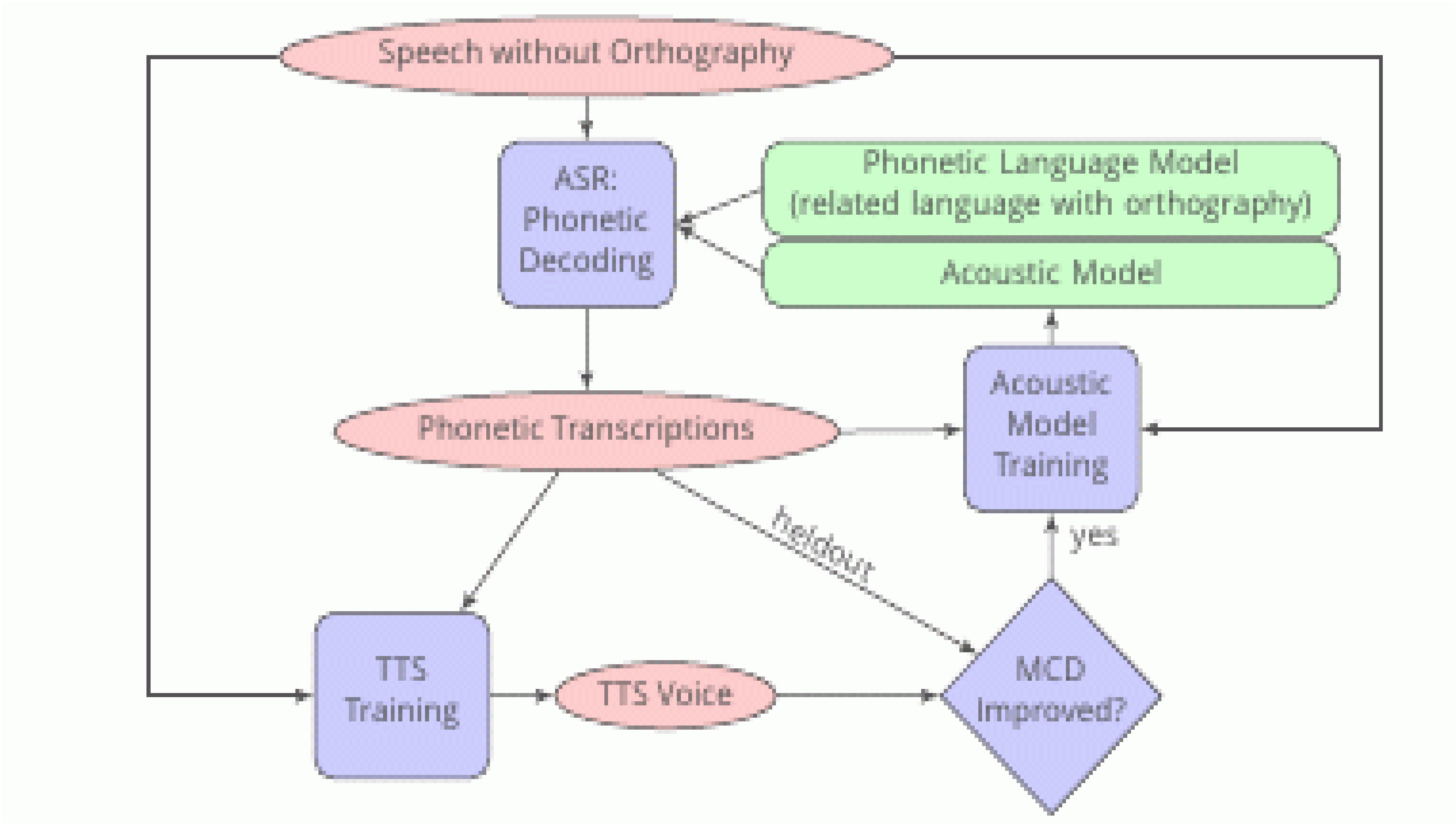
Deriving Writing



Cross Lingual Phonetic Labeling

- *For German audio*
  - *AM: English (WSJ)*
  - *LM: English*
  - *Example:*
- *For English audio*
  - *AM: Indic (IIIT)*
  - *LM: German*
  - *Example:*

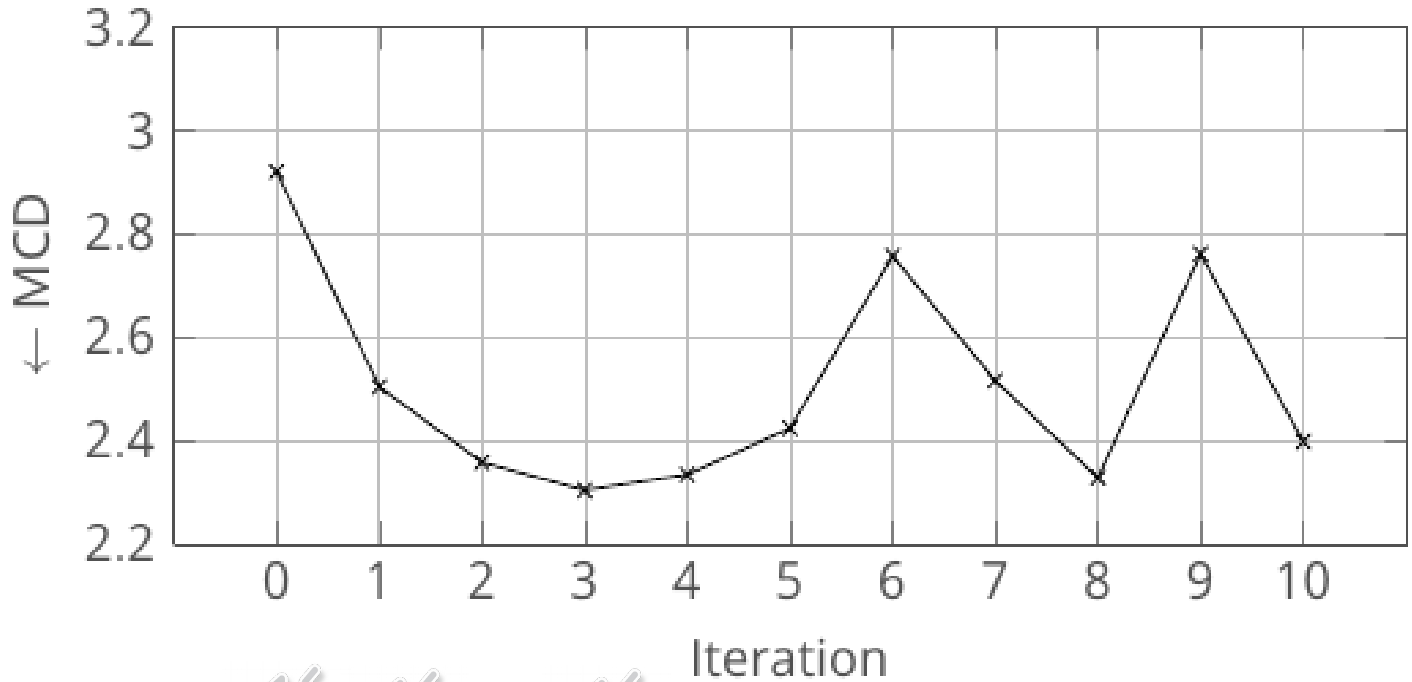
Iterative Decoding



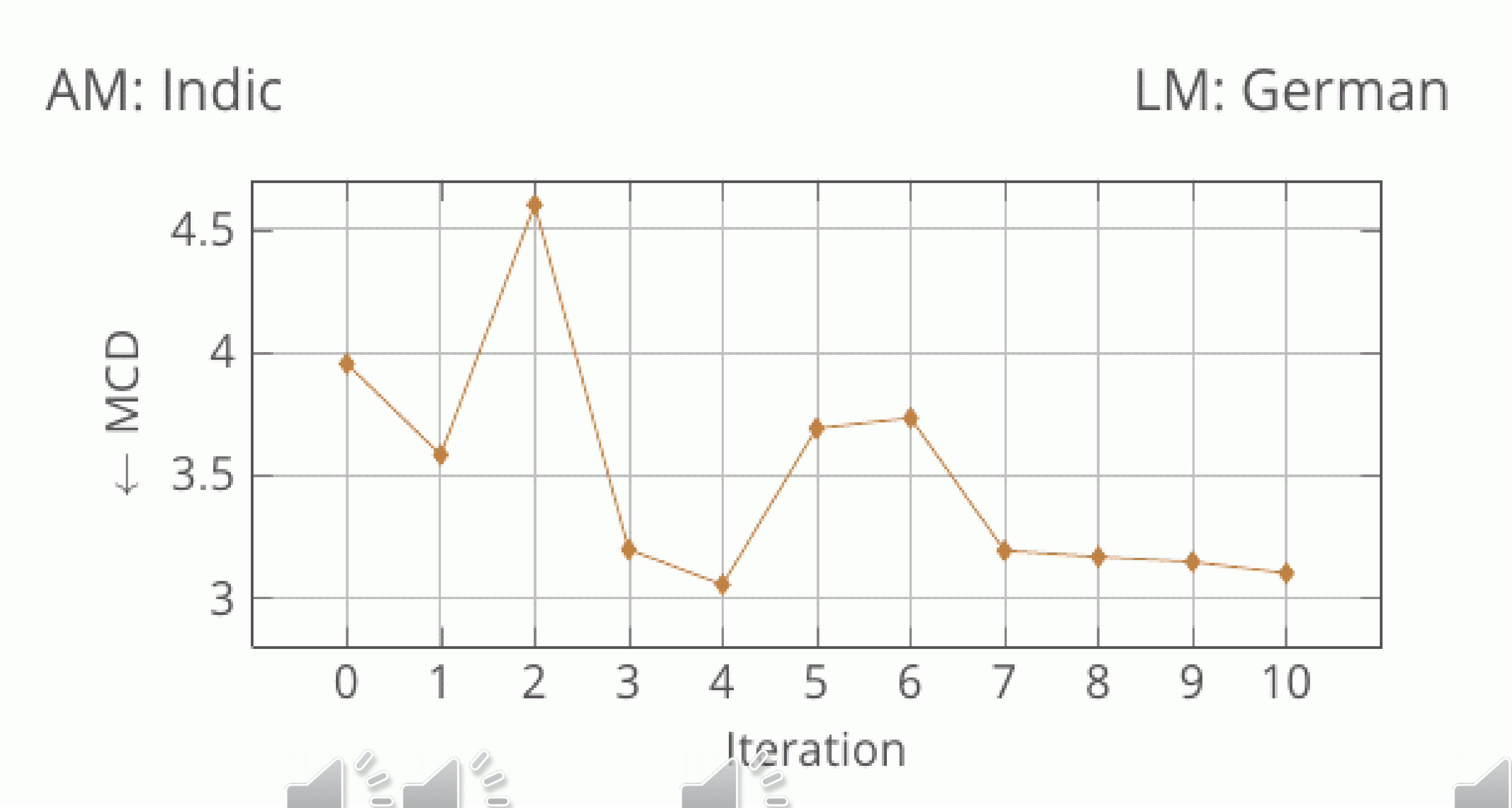
Iterative Decoding: German

AM: English (WSJ)

LM: English



Iterative Decoding: English



Find better Phonetic Units

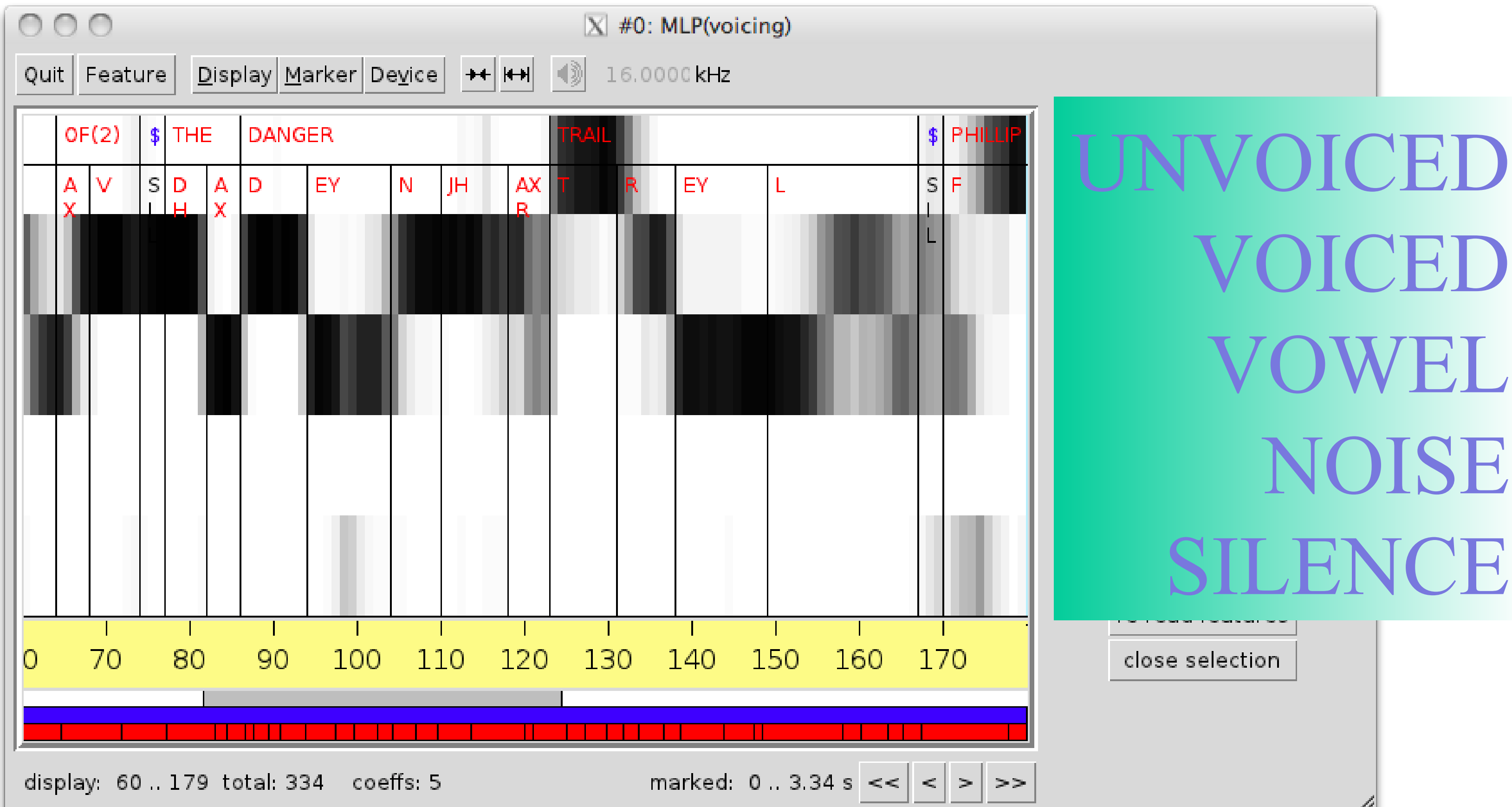
- Segment with cross lingual phonetic ASR
- Label data with Articulatory Features
  - (IPA phonetic features)
- Re-cluster with AFs

Articulatory Features (Metze)

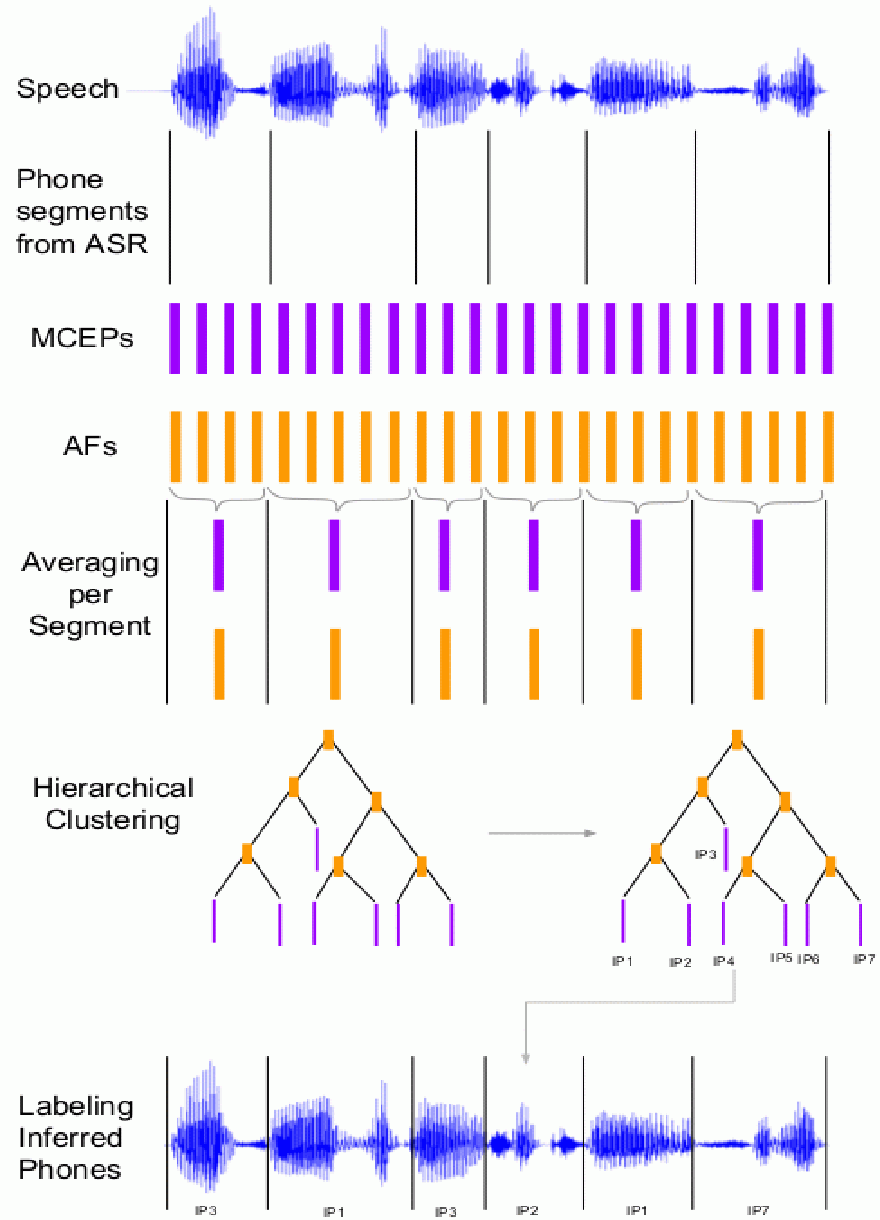
- 26 streams of AFs
- Train Neural Networks to predict them
  - Will work on unlabeled data
- Train on WSJ (Large amount **English** data)

ASR: "Articulatory" Features

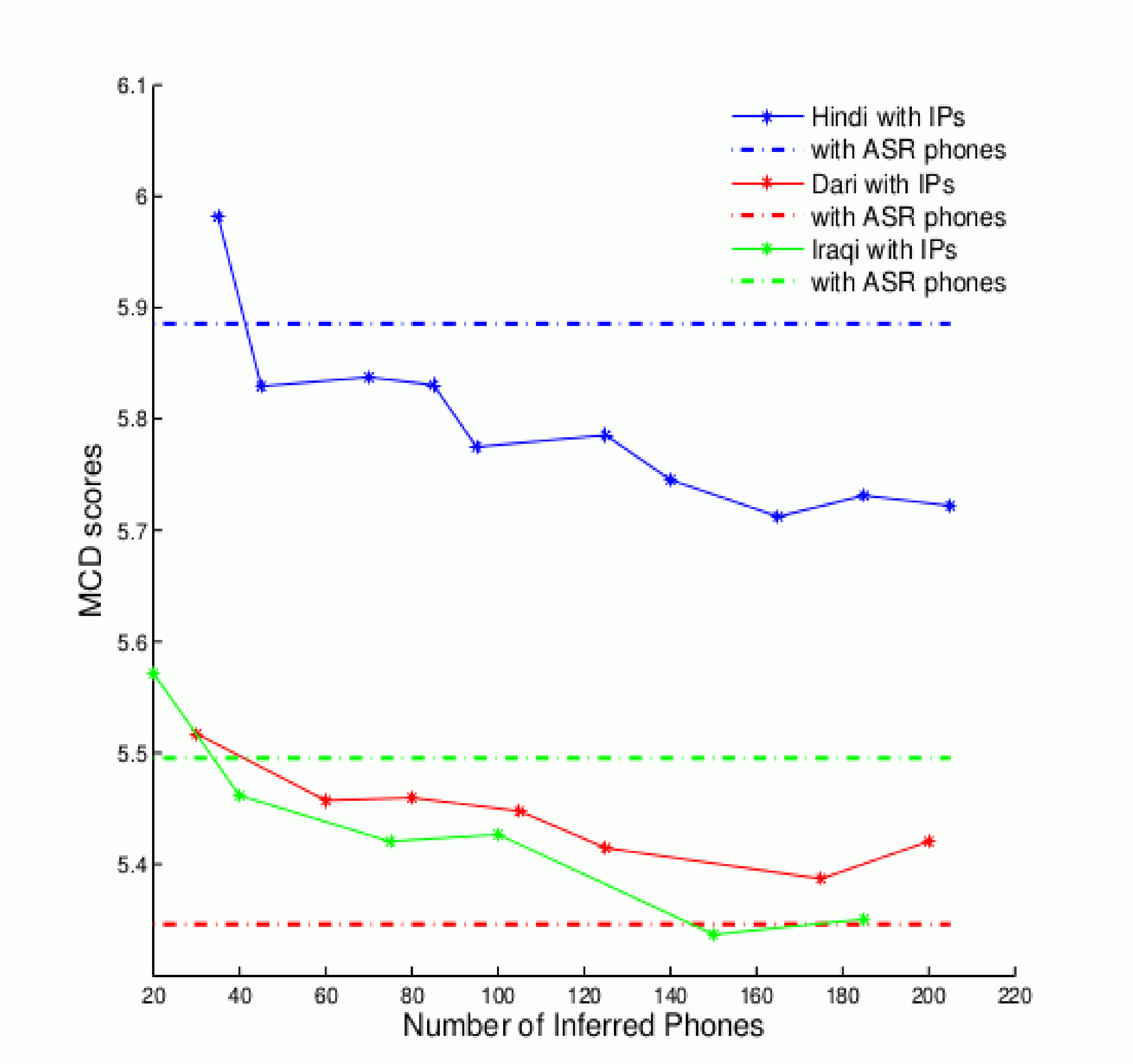
- *These seem to discriminate better*



Cluster New "Inferred Phones"



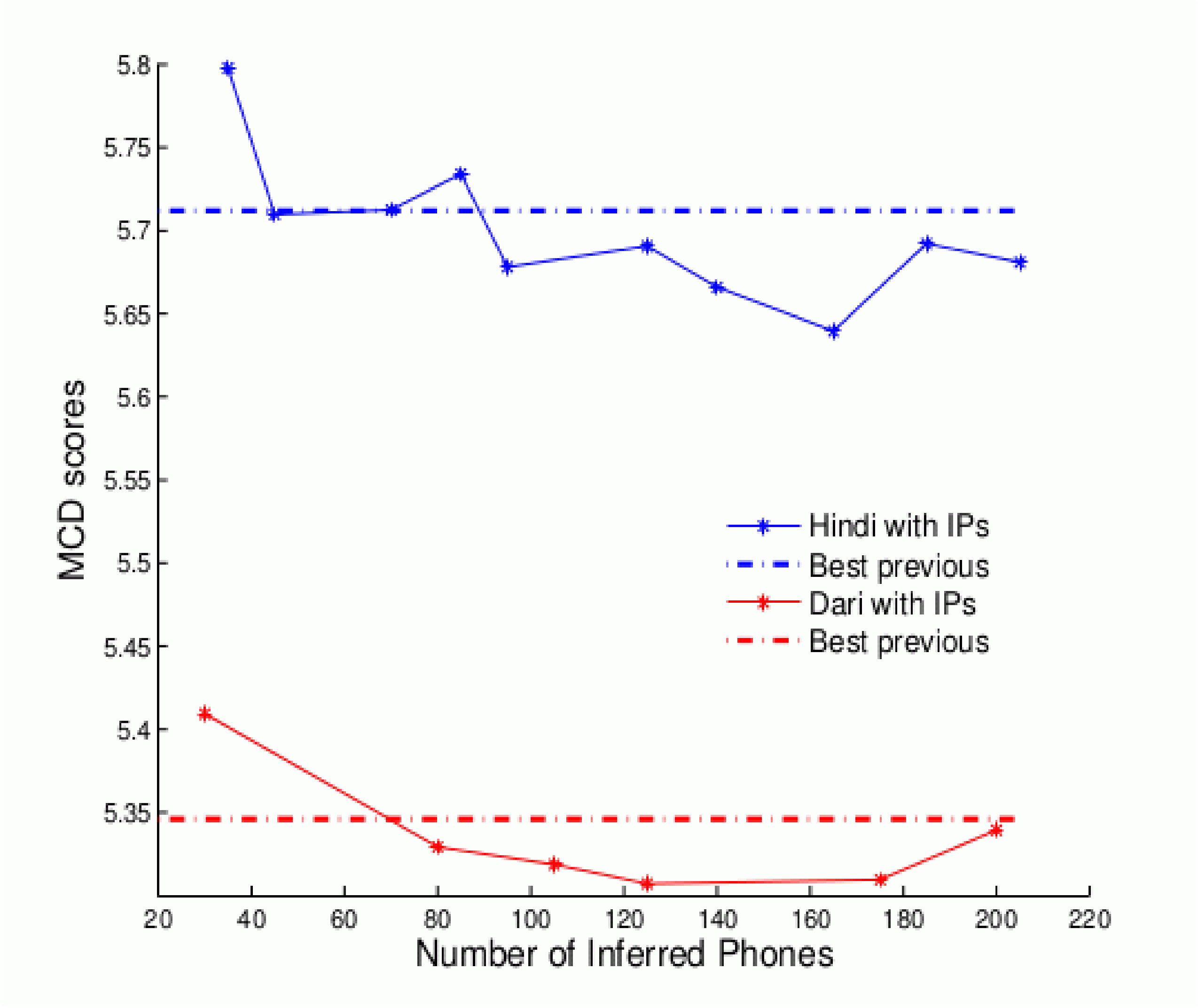
Synthesis with IPs



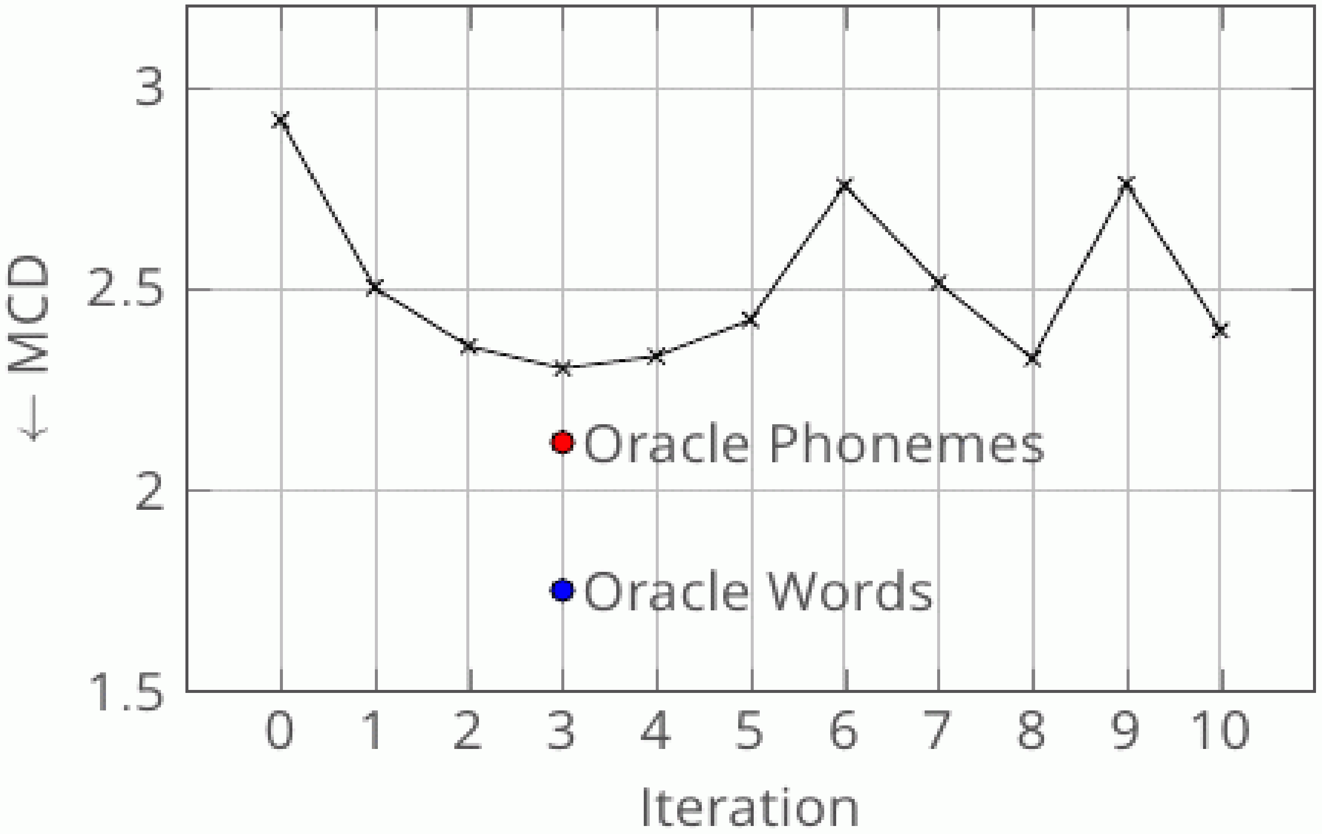
IP are just symbols

- *IPs don't mean anything*
  - *But we have AF data for each IP*
  - *Calculate mean AF value for each IP type*
    - *Voicing, Place of articulation ...*
  - *IP type plus mean/var AFs*

Synthesis with IP and AFs



German (Oracle)



Need to find "words"

- From phone streams to words
  - Phonetic variation
  - No boundaries
- Basic search space
  - Syllable definitions (lower bound)
  - SPAM (Accent Groups) (upper bound)
  - Deriving words (e.g Goldwater et al)

Other phenomena

- But its not just phonemes and intonation
  - Stress (and stress shifting)
  - Tones (and tone sondhi)
  - Syllable/Stress timing
  - Co-articulation
  - Others?
- [phrasing, part of speech, and intonation]
- MCD might not be sensitive enough for these
  - Other objective (and subjective measures)

But Wait ...

- Method to derive new “writing” system
- It is sufficient to represent speech
- But who is going to write it?

Speech to Speech Translation

- From high resource language
  - To low resource language
- Conventional S2S systems
  - ASR -> text -> MT -> text -> TTS
- Proposed S2S system
  - ASR -> derived text -> MT -> text -> TTS

Audio Speech Translations

- From audio in target language to text in another:
  - Low resources language (audio only)
  - Transcription in high resource language (text only)
- For example
  - Audio in Shanghaiese, Translation/Transcription in Mandarin
  - Audio in Konkani, Translation/Transcription in Hindi
  - Audio in Iraqi Dialect, Translation/Transcription in MSA
- How to collect such data
  - Find bilingual speakers
  - Prompt in high resource language
  - Record in target language

Collecting Translation Data

- Translated language not same as native language
- Words (influenced by English) (Telugu)
  - “doctor” → “Vaidhyudu”
  - “parking validation” → “???”
  - “brother” → “Older/younger brother”
- Prompt semantics might changes
  - Answer to “Are you in our system?”
  - Unnanu/Lenu (for “yes”/”no”)
  - Answer to “Do you have a pen?”
  - Undi/Ledu (for “yes”/”no”)

Audio Speech Translations

- Can’t easily collect enough data
  - Use existing parallel data and pretend one is unwritten
  - But most parallel data is text to text
- Let’s pretend English is a poorly written language

Audio Speech Translations

- Spanish -> English translation
  - But we need audio for English
  - 400K parallel **text** en-es (Europarl)
- Generate English Audio
  - Not from speakers (they didn't want to do it)
  - Synthesize English text with 8 different voices
  - Speech in English, Text in Spanish
- Use "universal" phone recognizer on English Speech
  - –Method 1: Actual Phones (derived from text)
  - –Method 2: ASR phones

English No Text

Table 1: Examples with raw phonemes

Original	I declare resumed the session of the European Parliament adjourned on ...
Method 1	ay d ih k l eh r r ih z uw m d dh ax s eh sh ax n aa v dh ax y uh r ax p iy ax n p aa r l ax m ax n t ax jh er n d aa n ...
Method 2	AY D IH K L EH R IY Z D UW IH NG DH IH S AE SH AH N AH V DH AE T Y AO R P IY AE N D P AA R T L IH M AE N D IH JH ER N D AA N ...

Phone to "words"

- Raw phones too different to Target (translation) words
  - Reordering may happen at phone level
- Can we cluster phone sequences as "words"
  - Syllable based
  - Frequent n-grams
  - Jointly optimize local and global subsequences
    - Sharon Goldwater (Princeton/Edinburgh)
- "words" do not need to be source language words
  - "of the" can be a word too (it is in other languages)

English: phones to syls

Table 2: Examples with naïve clustering

Method 1	ay d_ih k_l_eh_r_r_ih z_uw m_d_dh_ax_s_eh sh_ax_n aa v_dh_ax_y uh r_ax_p iy ax n_p_aa_r_l ax m_ax n_t_ax_jh er n_d_aa_n_f_r ...
Method 2	AY D_IH K_L_EH_R IY Z_D_UW IH NG_DH_IH_S AE SH_AH_N AH V_DH_AE_T_Y AO R_P_IY AE N_D_P_AA_R_T_L IH M_AE N_D_IH_JH ER N_D_AA_N_F_R ...

English: phones to ngrams

Table 3: Examples with most-frequent-ngrams clustering

Method 1	ay_d ih_k_l_eh_r r ih_z uw m d dh_ax_s_eh_sh_ax_n aa_v_dh_ax y_uh_r_ax_p_iy_ax_n p_aa_r_l_ax_m_ax_n_t_ax_jh er n_d aa_n …
Method 2	AY_D IH_K_L EH_R_IY_Z D UW_IH_NG DH_IH_S AE_SH AH_N AH_V_DH_AE_T Y_AO_R P_IY AE_N_D P_AA_R_T L IH_M_AE_N_D IH_JH ER_N_D AA_N …

English: phones to Goldwater

Table 4: Examples with Goldwater clustering

Method 1	aydihklehr rihzuwmddhaxseh shaxnaav dhaxyuhraxpiyaxn paarlaxmaxnt axjh-ernd aanfraydiy …
Method 2	AYDIHKL EHRIYZ DUWI-HNGDHIHS AESHAHNAHV DHAETYAORPIY AEND PAARTLIHM AEND IHJHERN D AAN …

English Audio → Spanish

Table 5: English-Spanish Results (BLEU)

	Words	Raw phonemes	Naïve syllables	Ngrams	Goldwater
Oracle	35.76				
Method 1		20.45	22.81	29.12	**31.92**
Method 2		13.81	13.78	18.46	**20.20**

Chinese audio → English

- 300K parallel sentences (FBIS)
  - Chinese synthesized with one voice
  - Recognized with ASR phone decoder

Table 7: Examples with different granularity

English gloss	an international audience
Word (hanzi)	国际 视听
Word (pinyin)	guójì shìtīng
Syllable (pinyin)	guó jì shì tīng
Phone (pinyin)	g uó j ì sh ì t ī ng
Goldwater (pinyin)	guójì shìtīng
Phone (ASR)	K L IH K S IY SH IY EY T S L IH M P
Goldwater (ASR)	KLIHKSIY SHIYEYT S LIHMP

Chinese Audio → English

Table 8: Mandarin-English Results (BLEU)

	Word	Syllable	Phone	Goldwater
Hanzi	29.05	27.27	N/A	N/A
Pinyin	28.98	26.78	14.29	26.80
Pinyin (toneless)	28.30	25.90	18.62	25.15
ASR	N/A	N/A	4.73	6.97

Spoken Dialog Systems

- Can we interpret unwritten languages
  - Audio -> phones -> “words”
  - Symbolic representation of speech
- SDS for unwritten languages:
  - SDS through translation
    - Konkani to Hindi S2S: + conventional SDS
  - SDS as end-to-end interpretation
    - Konkani to symbolic: + classifier for interpretation

Speech as Speech

- But speech is speech not text
  - What about conversational speech
  - Laughs, back channels, hesitations etc
  - Do not have good textual representation
  - Larger chunks allow translation/interpretation

"Text" for Unwritten Languages

- Phonetic representation from acoustics
  - Cross lingual, phonetic discovery
- Word representation from phonetic string
  - Larger chunks allow translation/interpretation
- Higher level linguistic function
  - Word classes (embeddings)
  - Phrasing
  - Intonation

Conclusions

- Unwritten languages are common
- They require interpretation
- Can create useful symbol representations
  - Phonetics, words, intonation, interpretation

- Let’s start processing speech as speech