Unwritten Languages as a Test Case for the Theory of Phonetic Universals

Mark Hasegawa-Johnson

University of Illinois

ISCSLP 2018

Outline

- speech-to-text and text-to-speech without the text
- speech2chat: generating text messages in an unwritten language
- inferring the phones and lexical tones
- image2speech: generating spoken descriptions of images

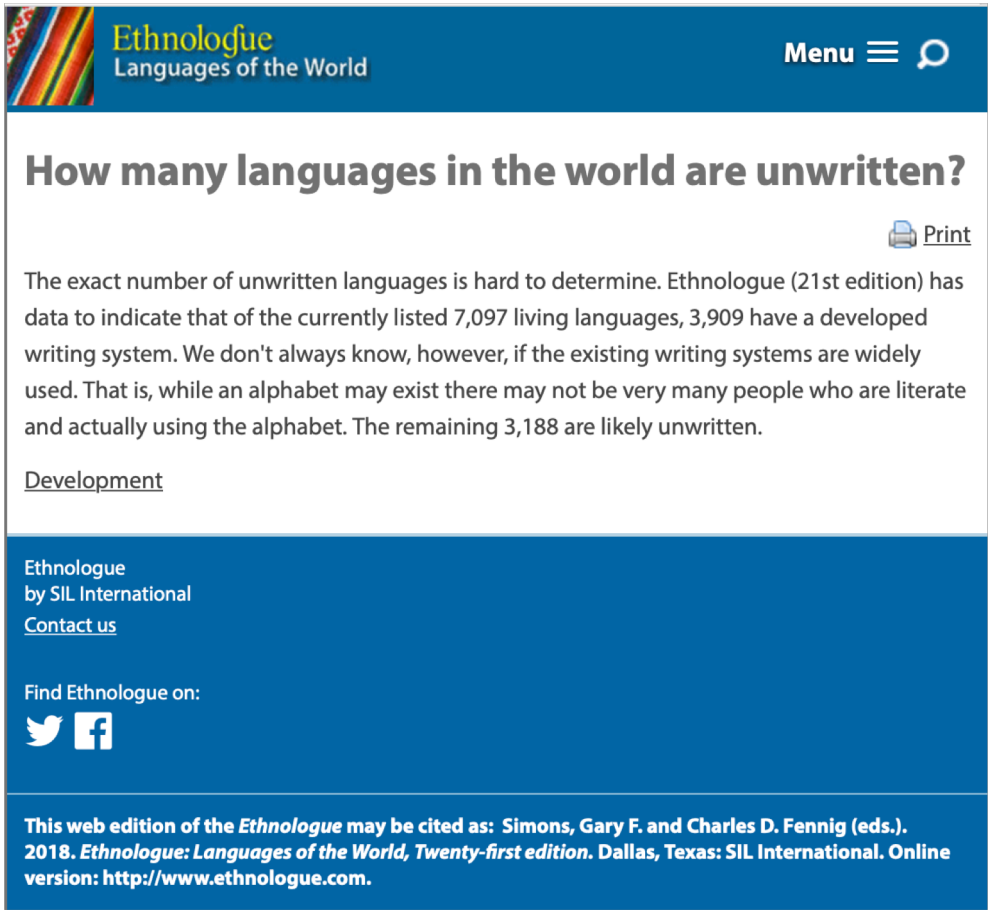
How many languages are there on Earth?

- Ethnologue: 7097 languages
  - Criterion: two different dialects whose native speakers are usually each unable to understand the other
- Graduate student in computer vision (2011): about 400
  - Criterion: there are 206 countries. Of those, about half are multilingual, with primary education in 2-20 languages each, with a mode of 3 languages.



How many languages are unwritten?

No writing system at all:
perhaps 3188 languages



Ethnologue
Languages of the World

Menu

How many languages in the world are unwritten?

Print

The exact number of unwritten languages is hard to determine. Ethnologue (21st edition) has data to indicate that of the currently listed 7,097 living languages, 3,909 have a developed writing system. We don't always know, however, if the existing writing systems are widely used. That is, while an alphabet may exist there may not be very many people who are literate and actually using the alphabet. The remaining 3,188 are likely unwritten.

Development

Ethnologue
by SIL International
Contact us

Find Ethnologue on:

This web edition of the *Ethnologue* may be cited as: Simons, Gary F. and Charles D. Fennig (eds.). 2018. *Ethnologue: Languages of the World, Twenty-first edition.* Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

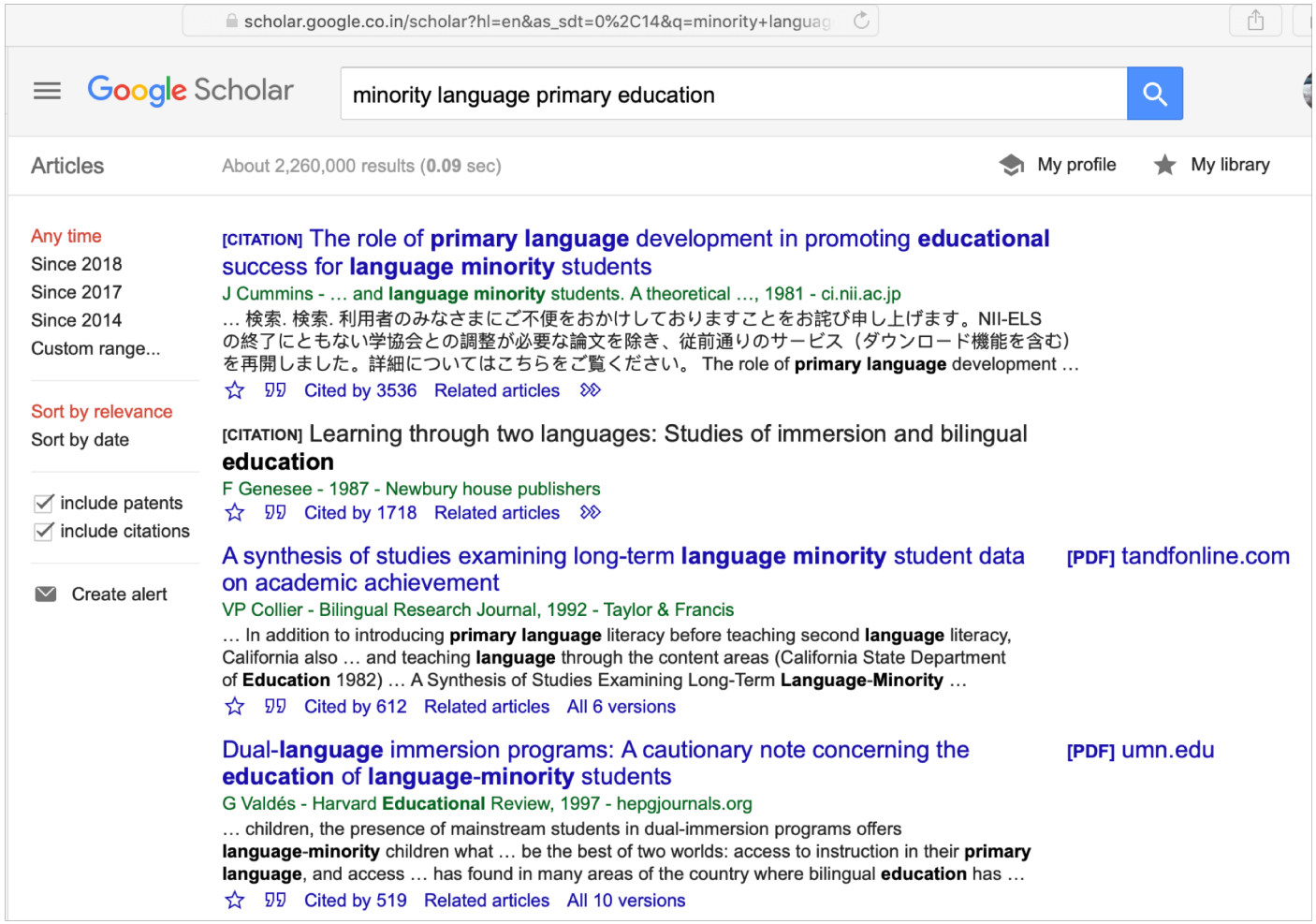
How many languages are unwritten?

Writing system exists but is not commonly used:

7097-400=6697

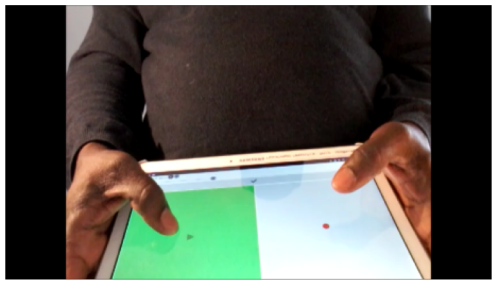
(Off-topic:

under exactly what circumstances should kids attend school in their native language vs. the majority language? United Nations: first 6 years of school.)



Example: Mboshi

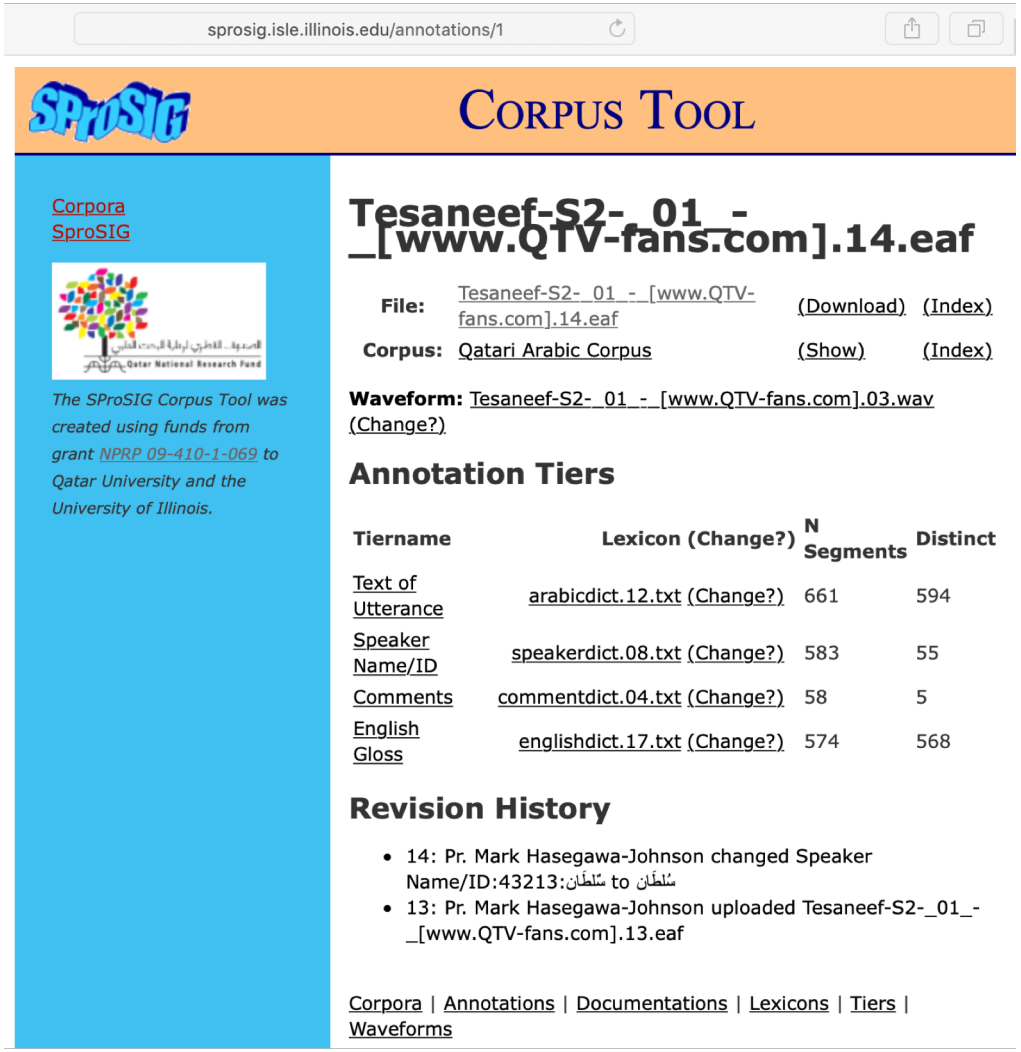
- **Mboshi** is spoken in Congo-Brazzaville – 110k native speakers
- *BULB Project Mboshi Corpus* built from translated reference sentences and from a Mboshi dictionary
  - G. Adda & al. *"Breaking the unwritten language barrier: The Bulb project"* in Proceedings of SLTU, Yogyakarta, Indonesia, 2016
  - collected with Lig_Aikuma mobile app: https://lig-aikuma.imag.fr
- (Human) French translations aligned to the speech utterances
- Language rarely written - linguists have defined a nonstandard grapheme form (close to language phonology) - forced alignments between speech & transcripts



táá ya bísí adí la kándzá lá iné	notre père a une grande paillotte
letsɛngɛ líibhámá l' ebhóyá	le sol durcit pendant la saison sèche

Example: Qatari Arabic

- Native language of ~1m people
- Gulf Arabic (macrolanguage): 6.8m native speakers
- Ethnologue calls Gulf Arabic a distinct language: ~33% words unique, ~33% declined uniquely.
- Native speakers consider it a dialect of standard Arabic. When sending text messages, use standard Arabic script with Qatari words & syntax.
- Standard Arabic ASR and TTS fail.



Example: Hokkien

- Native language of about 1.2m people in Singapore
- Southern Min (macrolanguage of Taiwanese, Hokkien, TeoChew): 28m speakers, 28th largest language in the world (Wikipedia)
- Used by characters in “Crazy Rich Asians” to emphasize that they are from Singapore. The book (by Kevin Kwan) contains more Hokkien than Mandarin.
- Latin orthographies exist, but are not taught in school.
- Native speakers pick and choose from Latin & Mandarin orthographies and/or make up spellings for
  - Person and Place names
  - Aphorisms in movie scripts and novels
  - Text messages



Michelle Yeoh and Calvin Wong on the set of “Crazy Rich Asians.” Image credit: Calvin Wong

Speech technology for unwritten languages: what does it mean?

It means you can…

- Play a song
- Call a friend
- Translate
- Send a text message
- Order a cat groomer

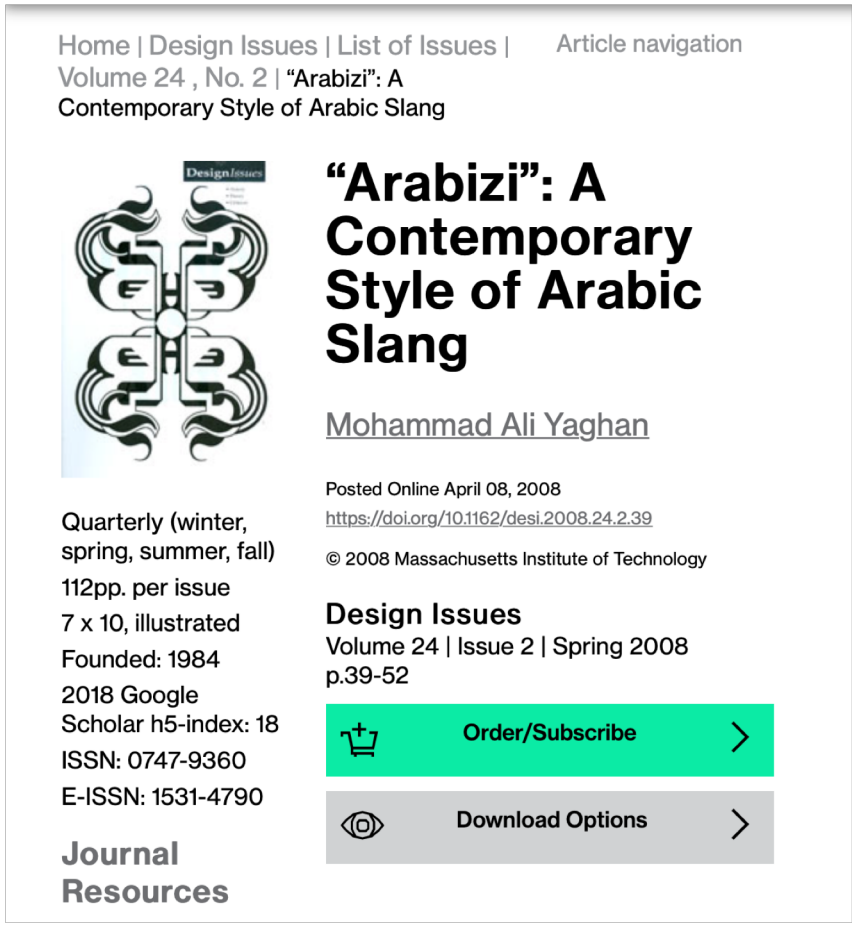
…using your own dialect, without having to switch to the majority dialect, even if your dialect is not mutually intelligible with the most similar written language.



Outline

- speech-to-text and text-to-speech without the text
- speech2chat: generating text messages in an unwritten language
- inferring the phones and lexical tones
- image2speech: generating spoken descriptions of images

Chat Alphabets



Home | Design Issues | List of Issues | Volume 24 , No. 2 | "Arabizi": A Contemporary Style of Arabic Slang

Article navigation

"Arabizi": A Contemporary Style of Arabic Slang

Mohammad Ali Yaghan

Posted Online April 08, 2008
https://doi.org/10.1162/desi.2008.24.2.39

© 2008 Massachusetts Institute of Technology

Quarterly (winter, spring, summer, fall)
112pp. per issue
7 x 10, illustrated
Founded: 1984
2018 Google Scholar h5-index: 18
ISSN: 0747-9360
E-ISSN: 1531-4790

Design Issues
Volume 24 | Issue 2 | Spring 2008
p.39-52

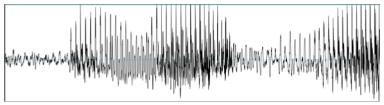
Order/Subscribe

Download Options

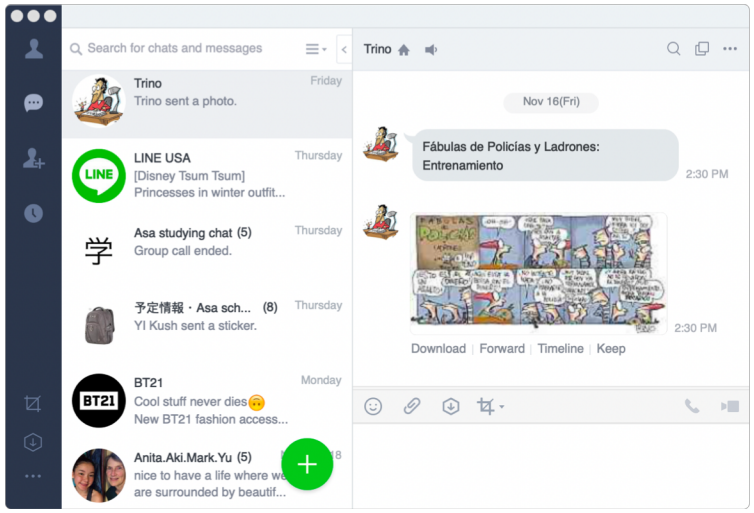
Journal Resources



speech2chat: generating text messages in a chat alphabet



Hey,
konshuumatsu
nanika shiyou,
jazz arushi,
luciferde



RAPID PHONETIC TRANSCRIPTION USING EVERYDAY LIFE NATURAL CHAT ALPHABET ORTHOGRAPHY FOR DIALECTAL ARABIC SPEECH RECOGNITION

Mohamed Elmahdy[⋆†] *Rainer Gruhn*[⋆] *Slim Abdennadher*[†] *Wolfgang Minker*[⋆]

ICASSP 2011:

Ask native speakers to transcribe audio using chat alphabet.

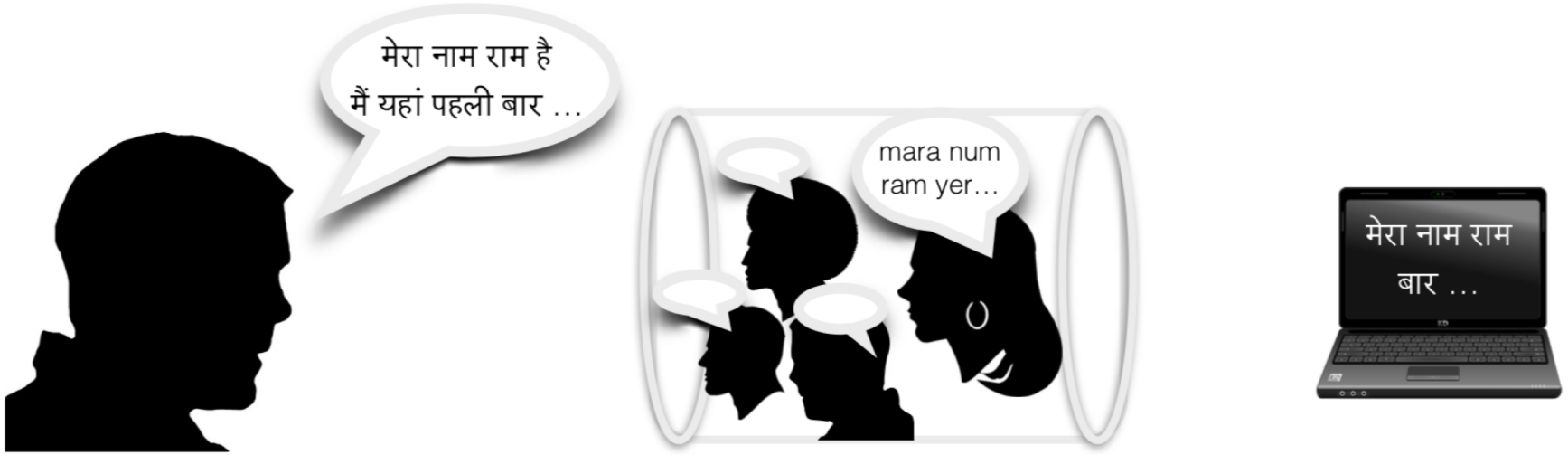
Why?

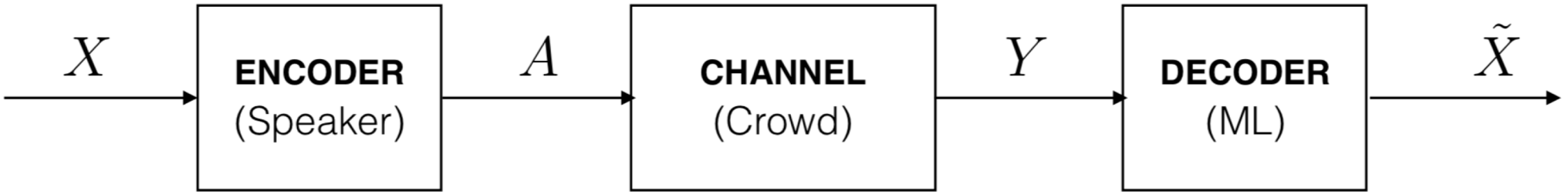
Allows transcription of Egyptian dialect words & syntactic forms that have no standard written form.

Arabic:	يوم الخميس ١ أكتوبر ٢٠١٠
English:	Thursday 1st of October 2010
IPA:	/yoːm ʔilxamiːs waːħid ʔuktoːbɑr ʔalfeːn wʕɑʃɑːrɑ/
ACA:	/yoom 2il7'amiis waa7ed 2uktoobar 2alfeen w3ashaara/
Arabic:	تذكرة درجة أولى
English:	First class ticket
IPA:	/tɑzkɑrɑ dɑrɑgɑ ʔuːla/
ACA:	/tazkara daraga 2uula/
Arabic:	بيض مسلوق
English:	Boiled eggs
IPA:	/beːḍ masluːʔ/
ACA:	/bee9' masluu2/

An even noisier proxy for chat alphabet transcripts: mismatched (non-native) transcripts

(http://www.isle.illinois.edu/sst/data/sbs_transcripts/index.html)





Tongan is a Polynesian language spoken mainly on the island of Tonga in the Pacific Ocean. What does Tongan sound like to people who don't know it? Are there systematic patterns in how non-native listeners perceive speech in Tongan? These are some of the questions we are exploring.

Please attempt this HIT only if you do not speak or understand Tongan. You will be asked to listen to **eight** very short audio clips (roughly 1 second long) in Tongan. Your task is to write down what you hear using English text that sounds roughly the same. Please provide English text for **all eight audio clips** to ensure that your HITs will be approved. If the clip contains only music, please write down **"[music]"** in the text box. If the clip is empty, please write **"[empty]"** in the text box. Please ensure that any comments you want to include in the text boxes (other than the text matching the audio) are enclosed within square brackets, "[]".

IMPORTANT: Please use English text that most closely matches what you hear in the speech (which would typically be 'nonsense words' in English). The example clip below will clarify this further.

IMPORTANT NOTE: Please make sure that, as far as possible, the English text you provide closely matches the audio. Common proper nouns (names of countries, etc.) appearing within the sound clips are expected to be identified correctly. If you have attempted these HITs previously, please note that we have some new quality checks in place now: HITs will be randomly checked and rejected if they do not comply with the task instructions.

The audio clips are extracted from Special Broadcasting Service (SBS) Australia radio podcasts that typically contain a mix of news, current affairs and other community information.

BONUS: For every 12 HITs that you successfully complete and that get approved, you will receive a bonus of 25 cents. Please note that the bonus is not prorated. For example, if you do 24 HITs or 35 HITs, your bonus will be 50 cents. **Bonuses will be approved within a week after the entire batch of HITs are completed.**

Before you begin the HIT, please enter a comma-separated list of all the languages you are familiar with and understand. **You only need to fill this in once and not for every HIT.**

List of languages goes here

Clip 1:



If the above player does not work, try this:

FLASH

Or, click here for the audio file.

Or, use this player if you would like the clip to play on loop until stopped:

FLASH

Text goes here

Clip 5:



If the above player does not work, try this:

FLASH

Or, click here for the audio file.

Or, use this player if you would like the clip to play on loop until stopped:

FLASH

Text goes here

Clip 2:



Clip 6:



Transcript of swahili_140826_356959-5 by Native Speaker

-da pamoja na . . .

Transcriptions Provided by People who Don't Speak Swahili

- e n d a p o m o n y a m A
- m e m b e r p a m o h y a
- e n d a c u m o y a h
- e n d a a m o y a a
- e n d a p a m o y a n a
- e n d a p a m o l y a
- e n d a p u h m o h y a n A
- e b a p u m u l i e u h
- e n n a p a m U y a n A
- d a p a m o j e r a

The basic idea: Non-Native Transcription ≈ Noisy Phonetic Transcription

English Nonsense Orthography	Implied IPA
enda	/ɛnda/
pamooya	/pʰəmuja/
baitable	/betəbl̩/

Models of Second-Language Speech Perception: Perceptual Assimilation Model (Best et al., 1988, 2009)

A listener's ability to distinguish two foreign phones, f_1 and f_2, is determined by the native phones e_1 and e_2 to which they are perceptually most similar:

$$e_1 = \text{argmin}_{e \in \mathcal{E}} D(e, f_1)$$

$$e_2 = \text{argmin}_{e \in \mathcal{E}} D(e, f_2)$$

where $\mathcal{E}$ means the native phone set, and $D(e, f)$ means the perceptual distance between native phoneme e and foreign phoneme f.

$$e_k = \text{argmin}_{e \in \mathcal{E}} D(e, f_k)$$

/ɮ/  /ɬ/ 

TC	Two-Category	$e_1 \neq e_2$
ABX discrimination accuracy: 95%		
Example: Zulu voiced vs. unvoiced lateral fricatives		

/ǁ/  /ǃ/ 

NA	Not Assimilable	$e_1 = \emptyset,\ e_2 = \emptyset$
ABX discrimination accuracy: 81–99%		
Example: Zulu clicks		

/kʰ/  /kʼ/ 

CG	Category Goodness	$e_1 = e_2$ but $D(e_1, f_1) \ll D(e_1, f_2)$
ABX discrimination accuracy: 88%		
Example: voiceless aspirated versus ejective velar stops		

/b/  /ɓ/ 

SC	Single Category	$e_1 = e_2$ and $D(e_1, f_1) \approx D(e_1, f_2)$
ABX discrimination accuracy: 65%		
Example: Zulu plosive vs. implosive bilabial stops		

Computing the Data Likelihood

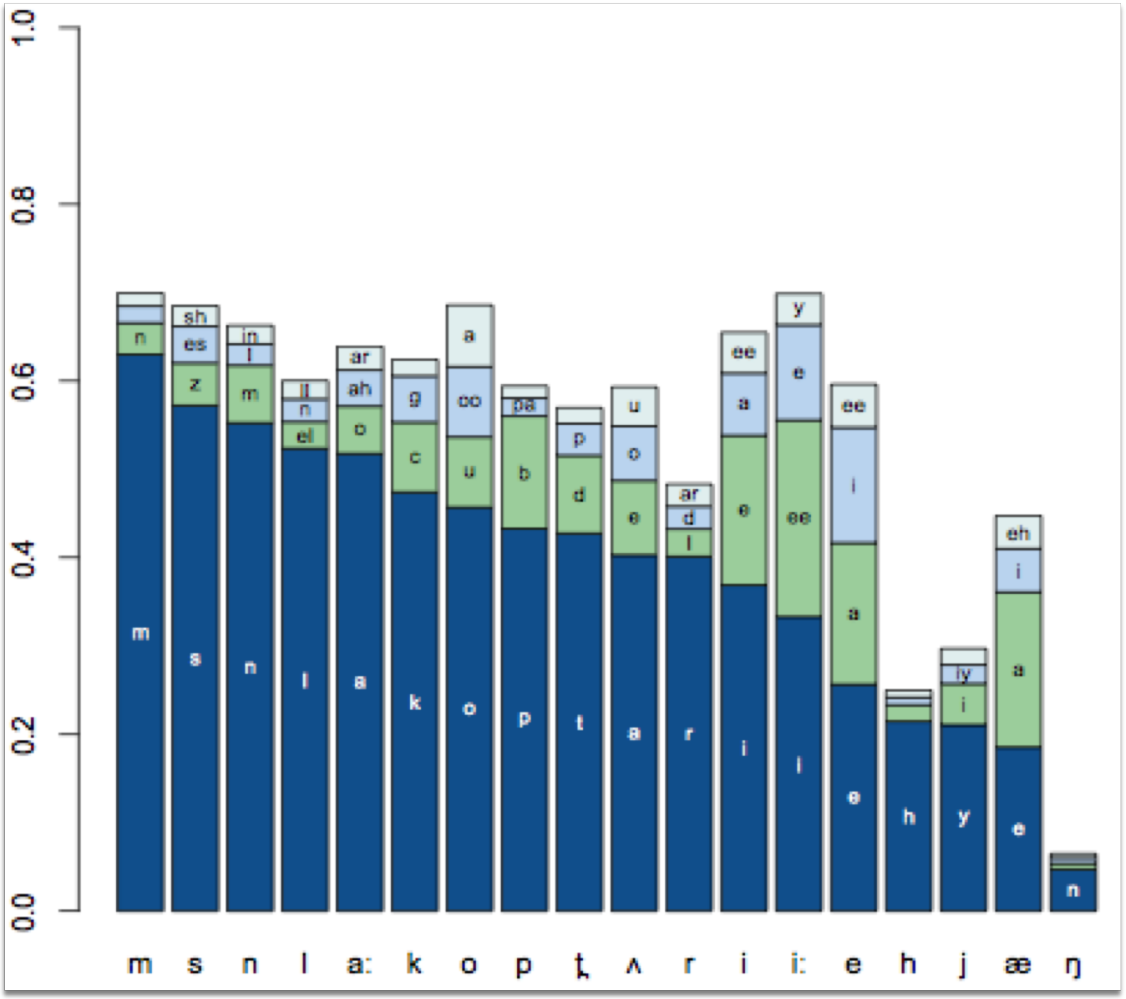
Suppose we've trained a finite state model of the mismatch channel; its states are s_t, its input alphabet is y_t (the spoken phone), its output alphabet is $x_{t\ell}$ (the written symbol). Given transition probabilities $a(s_t|s_{t-1})$ and observation probabilities $b(x_t|s_t)$ we can compute the likelihood of training data as

$$p(\vec{x},\vec{y}) = \sum_{\vec{s}\in\vec{y}} \prod_{\ell=1}^{L} \prod_{t=1}^{T} b(x_{t\ell}|x_t)a(s_t|s_{t-1})$$

. . . and the posterior probability of an $s_{t-1} \rightarrow s_t$ state transition as

$$\gamma(s_t, s_{t-1}|\vec{x},\vec{y}) = \frac{\prod_{\ell=1}^{L}\prod_{t=1}^{T} b(x_{t\ell}|x_t)a(s_t|s_{t-1})}{p(\vec{x},\vec{y})}$$

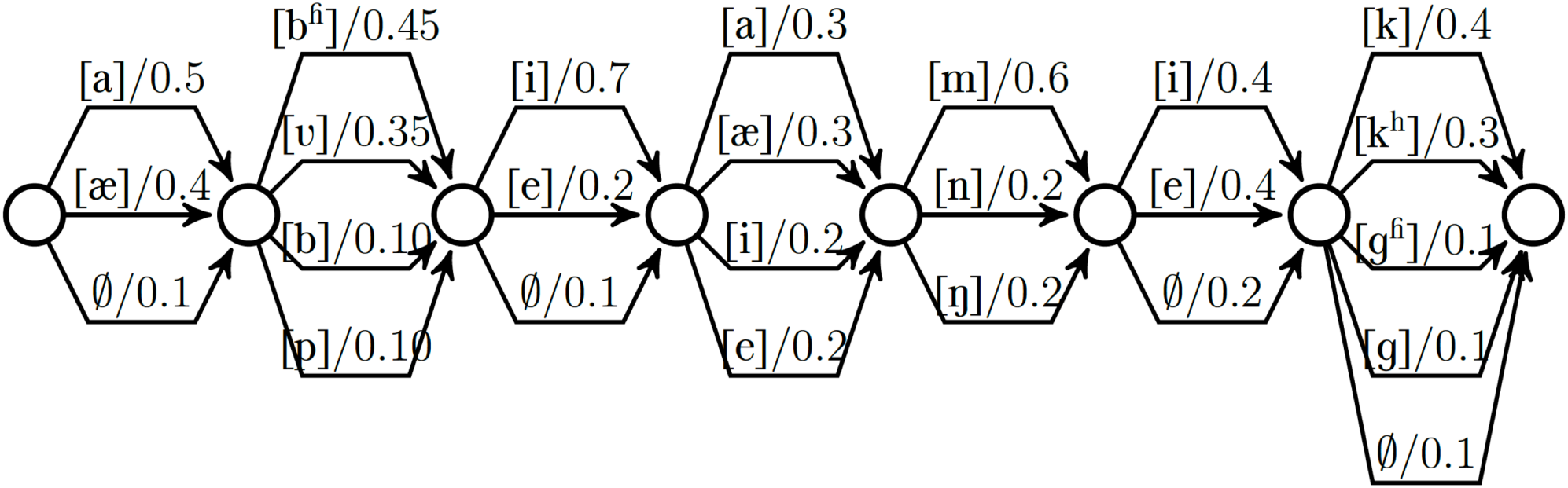
A Trained Model of Non-Native Speech Perception: Unigram Statistics (Hindi→English)



Jyothi & Hasegawa-Johnson, AAAI 2015

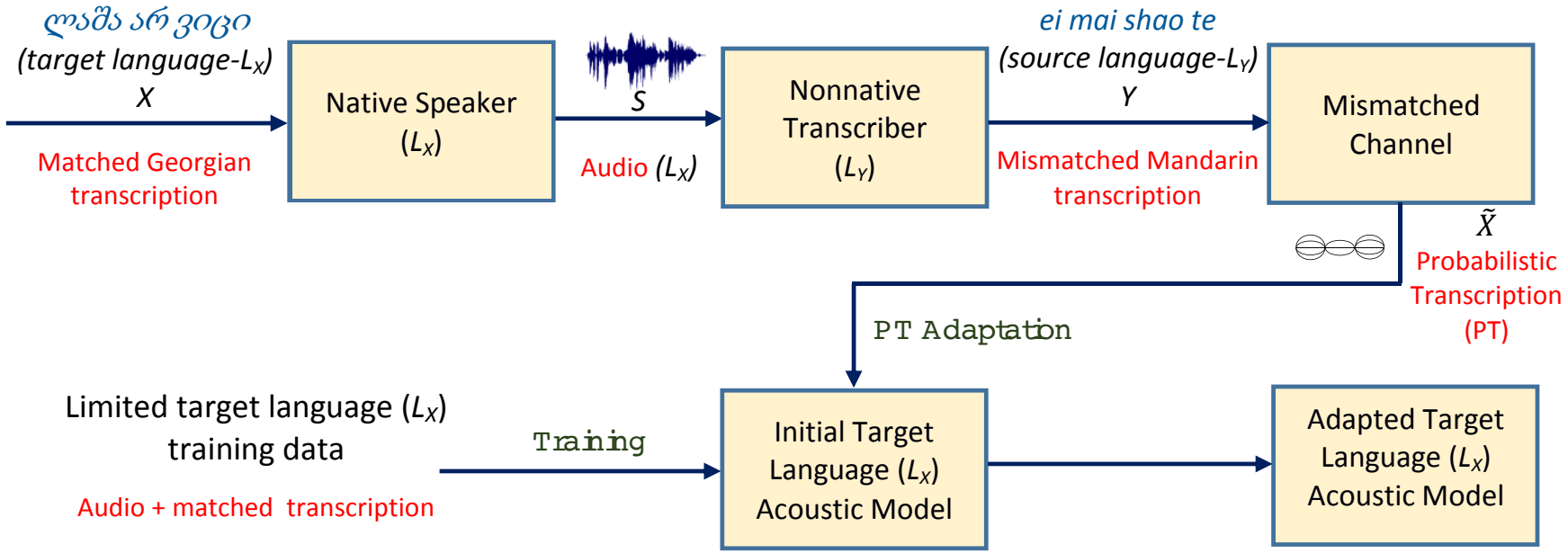
Probabilistic Transcript

Several different mismatched transcripts are time-aligned, then converted to a “Probabilistic Transcript” (a probability mass function over the set of possible transcriptions).



ASR Trained Using Probabilistic Transcripts

- Native speaker of under-resourced language (e.g., Georgian) records some audio
- Non-speaker of under-resourced language (Georgian, Hokkien, etc.) transcribes it
- Machine converts this to a “probabilistic transcription” (a lattice), then trains an ASR



PT-Adapted ASR has lower error rate than (1) cross-language ASR, (2) human transcribers

Phone error rates (lower is better):

Black=ASR error rate, Red=human error rate

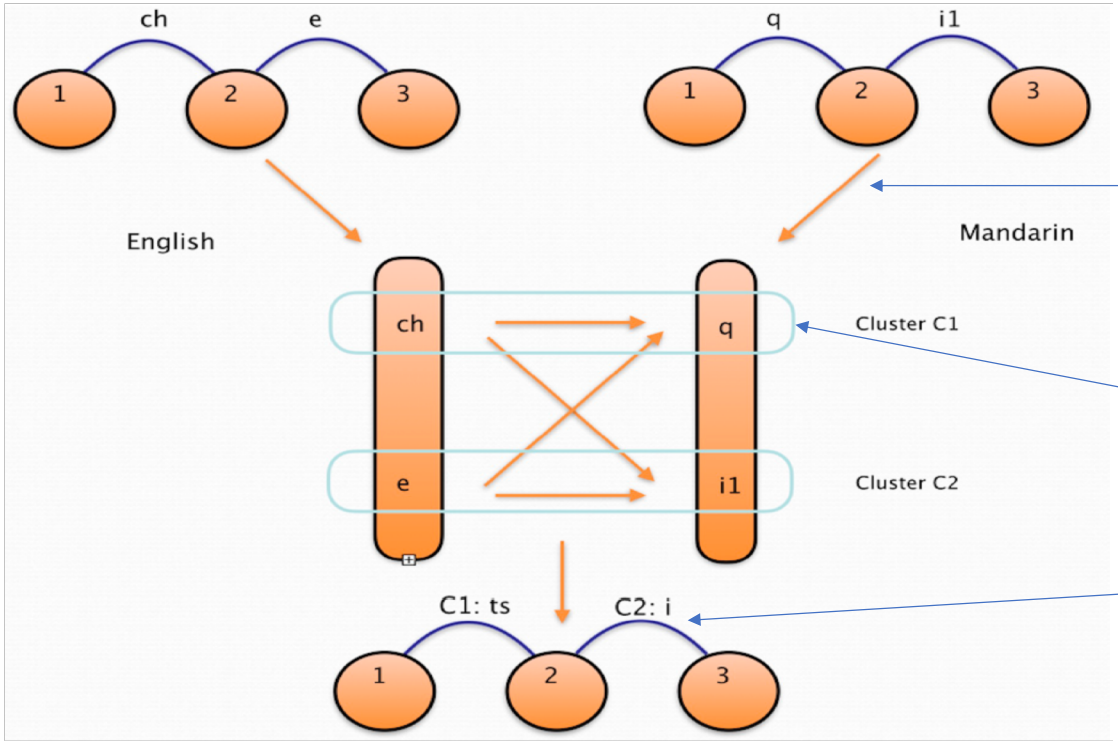
Training	Hungarian	Mandarin	Swahili
Cross-Language GMM	67 (%)	69	65
Cross-Language DNN	64	61	60
PT-Adapted GMM	**58**	57	49
PT-Adapted DNN	**58**	**53**	49
Humans	64 (%)	71	**48**

Outline

- speech-to-text and text-to-speech without the text
- speech2chat: generating text messages in an unwritten language
- inferring the phones and lexical tones
- image2speech: generating spoken descriptions of images

inferring the phones of Hokkien by clustering Mandarin & English mismatched transcripts

(Wenda Chen et al., Interspeech 2017)



Cluster Learning

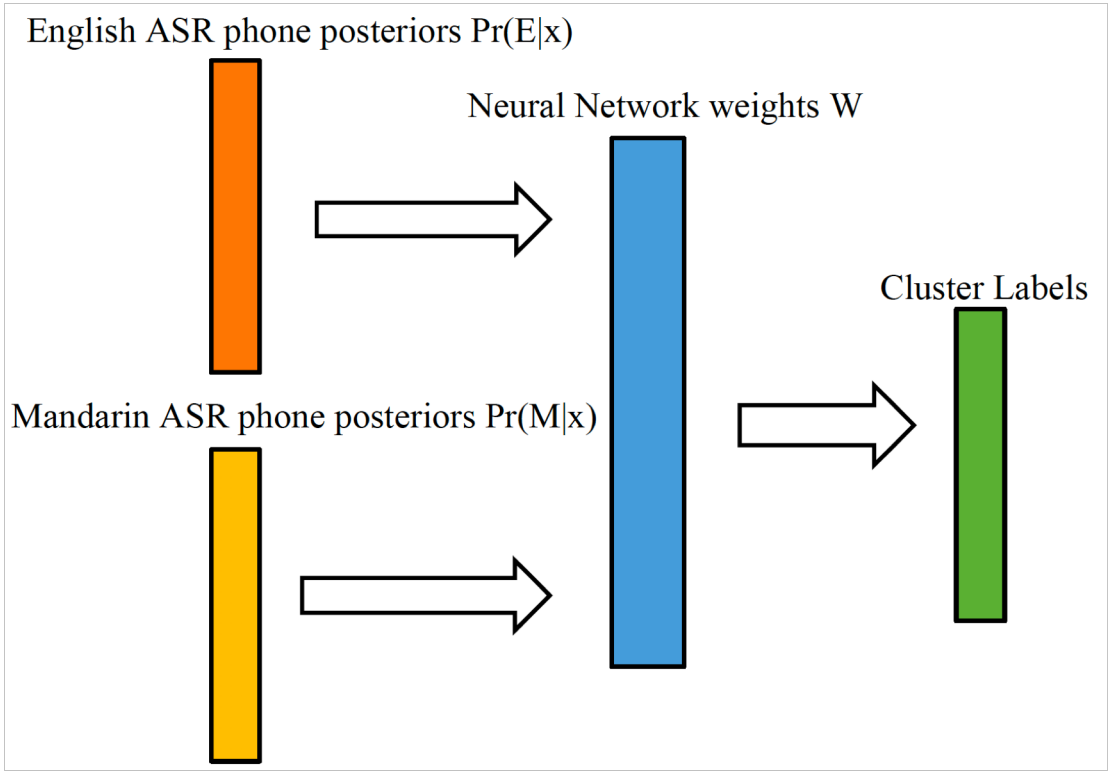
- English-speaker transcriptions and Mandarin-speaker transcriptions are time-aligned.
- Alignment edges are clustered to defined “phone set” of the under-resourced language.
- **Cluster Learning**: DNN is trained to output these cluster symbols.

inferring the phones of Hokkien by clustering Mandarin & English mismatched transcripts

(Wenda Chen et al., Interspeech 2017)

Cluster Re-Training

DNNs are pre-trained in English and Mandarin, then a new softmax output layer is trained to combine them, in order to estimate phone probabilities in the under-resourced language.



inferring the phones of Hokkien by clustering Mandarin & English mismatched transcripts

(Wenda Chen et al., Interspeech 2017)

Phone error rates (lower is better)

Algorithm	Vietnamese	Hokkien
DNN trained on 40 hours of English	84%	83%
DNN trained on 40 hours of Mandarin	78%	73%
GMM, MAP-adapted to a Probabilistic Transcription from Mismatched Crowdsourcing	77%	73%
Multilingual DNN, re-trained using unlabeled data in the target language	73%	**67**%
Cluster Learning		**67**%
Cluster Re-Training	**70**%	**67**%

inferring lexical tones from non-native transcription

1. Collect confusion matrix, $P\left(t_j^L \middle| t_i^K\right)$
   - t_j^L = j'th lexical tone in language L, e.g., $t_1^m =$ Mandarin tone 1
   - t_i^K = i'th lexical tone in language K, e.g., $t_3^c =$ Cantonese tone 3
2. Model the confusion matrix using multi-dimensional scaling: find two-dimensional vectors x_j^L so that

$$P\left(t_j^L \middle| t_i^K\right) \approx e^{-\left\| x_j^L - x_i^K \right\|}$$

inferring lexical tones from non-native transcription: DATA

Mismatched crowdsourcing

(Wenda Chen et al., Procedia Computer Science 2016)

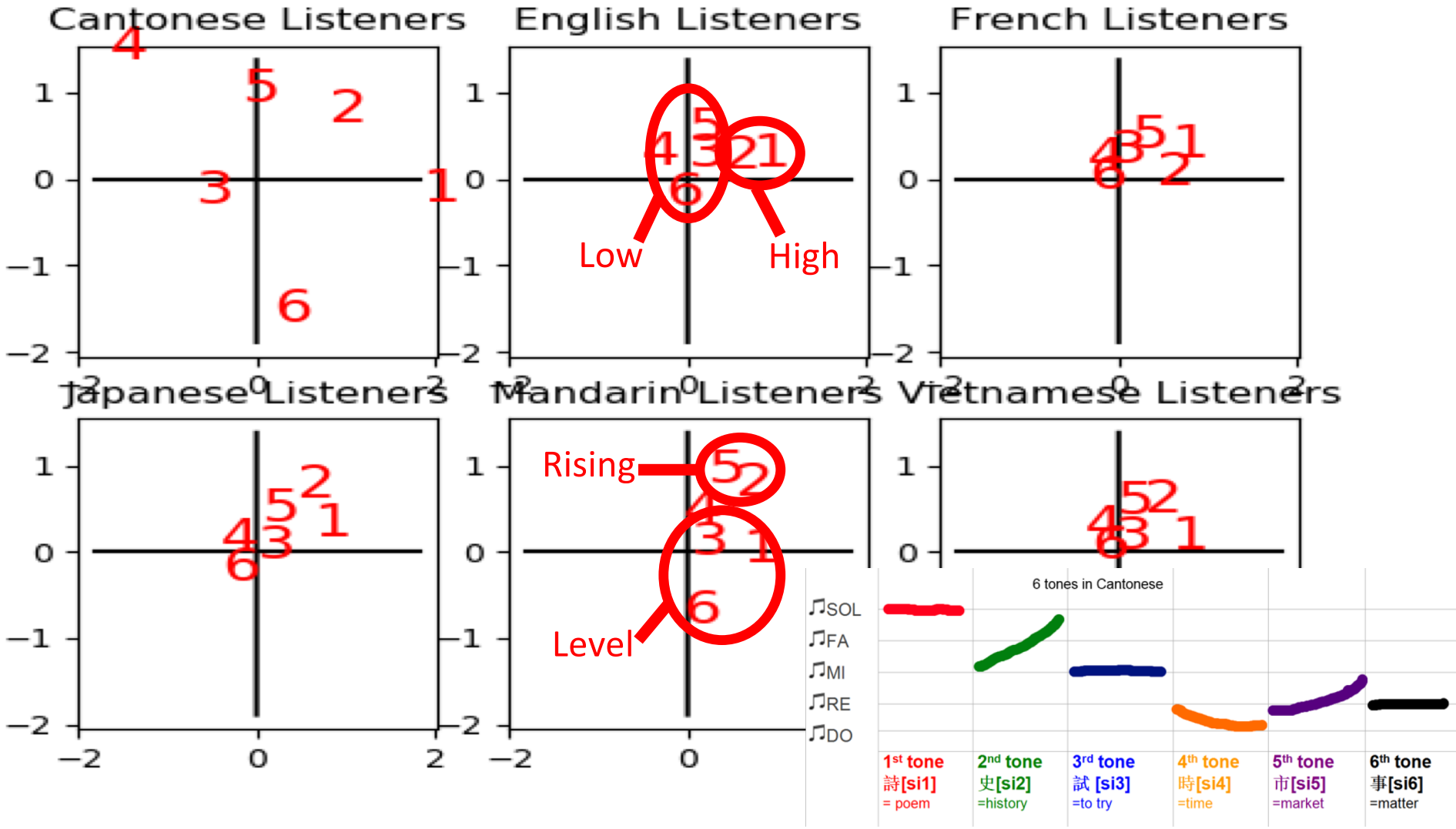
- Vietnamese, Cantonese speech
- Vietnamese, Cantonese, Mandarin listeners, using their own tone numbers

Non-speakers trained to use the correct labels

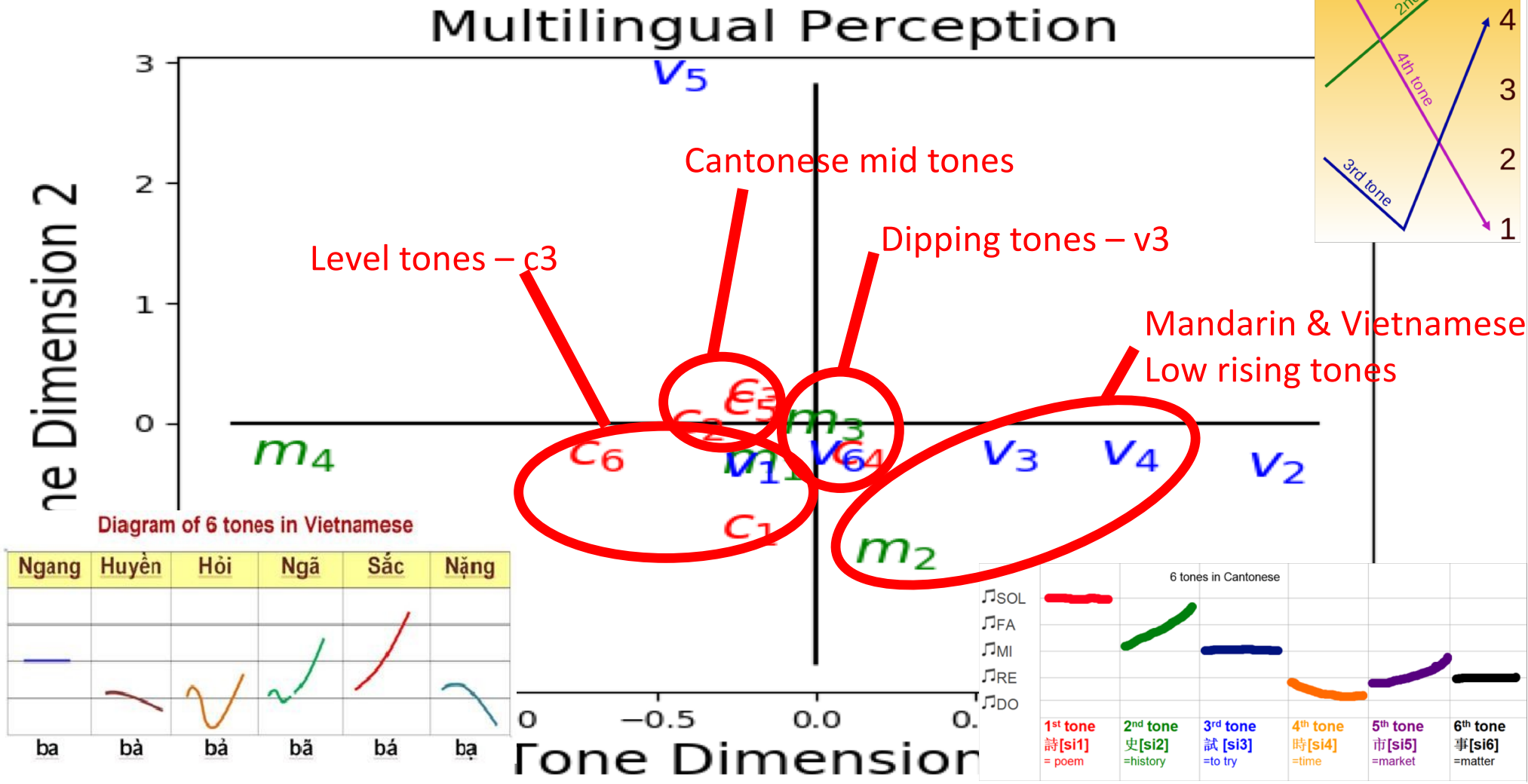
(Nancy Chen et al., not-yet-published data)

- Cantonese speech
- Cantonese, Mandarin, Vietnamese, Japanese, American, French listeners

Cantonese tones



All tones, all listeners grouped together



Outline

- Speech technology for unwritten languages: what does it mean?
- speech2chat: generating text messages in an unwritten language
- inferring the phones and lexical tones
- image2speech: generating spoken descriptions of images

The img2txt task

(a.k.a. image captioning)

Framing image description as a ranking task: data, models and evaluation metrics

Micah Hodosh Peter Young Julia Hockenmaier

Annotation Demo

What follows is a sample of the results of image annotation of the final lexical and semantic kernels of

M. Hodosh, P. Young and J. Hockenmaier (2013) "Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics", Journal of Artificial Intelligence Research, Volume 47, pages 853-899

http://www.jair.org/papers/paper3994.html

Each query image is shown with the top 5 caption results from our test set.
If it appears, the original caption is bolded

Query:



Top Results:

A man is standing at the shoreline on the beach with 4 dogs .
A man playing with his dog on a beach
a dog on the beach .
A man kneels in the wet sand at a beach near a dog .
A brown dog leaps into the air next to a black dog on a sandy bay .

image2speech = img2txt + text2speech - text

- If there is no writing system, then speech is the only communication tool, but:

Must one speak in the majority dialect to be understood by one's cell phone?

- Task Definition: Can we develop speech technology in a dialect that is almost never written down in any standardized text format?

Definition: An **image2speech** algorithm is an algorithm that observes an image, and generates a spoken description of the image, without requiring that the language of the description has any standardized text format.

Datasets: flickr8k (American English, Natural)

- Crowd workers on Mechanical Turk recorded the texts of flickr8k**.**
- 40k spoken captions are available online:
  - https://groups.csail.mit.edu/sls/downloads/
  - D. Harwath and J. Glass, *"Deep multimodal semantic embeddings for speech and images"* in IEEE ASRU, Scottsdale, Arizona, USA, December 2015



-A brown and white dog is running through the snow

-A dog is running in the snow

-A dog running through snow

-A white and brown dog is running through a snow covered field

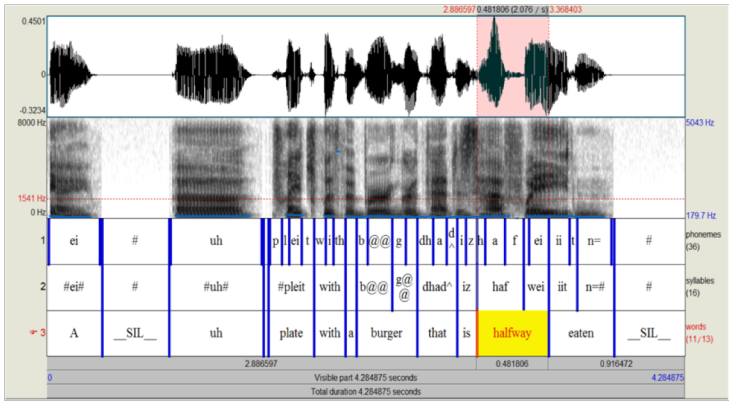
-The white and brown dog is running over the surface of the snow

Datasets: SpeechCOCO (UK English, Synthetic)

- Augmentation of MSCOCO with (Voxygen) TTS Speech
  - W. Havard, L. Besacier & O. Rosec, *"SPEECH-COCO: 600k Visually Grounded Spoken Captions Aligned to MSCOCO"*, GLU Workshop of Interspeech, Stockholm, Sweden, 2017
  - https://persyval-platform.imag.fr/perscido/web/DS80/detaildataset
- Disfluencies and speed perturbation to add variability
- WAV paired with JSON containing timecode of each word/syllable/phoneme



+



A plate with a burger that is halfway eaten.

image2speech system components

Image Representation = VGG16

Machine Translation = XNMT

Speech Synthesis = ClusterGen

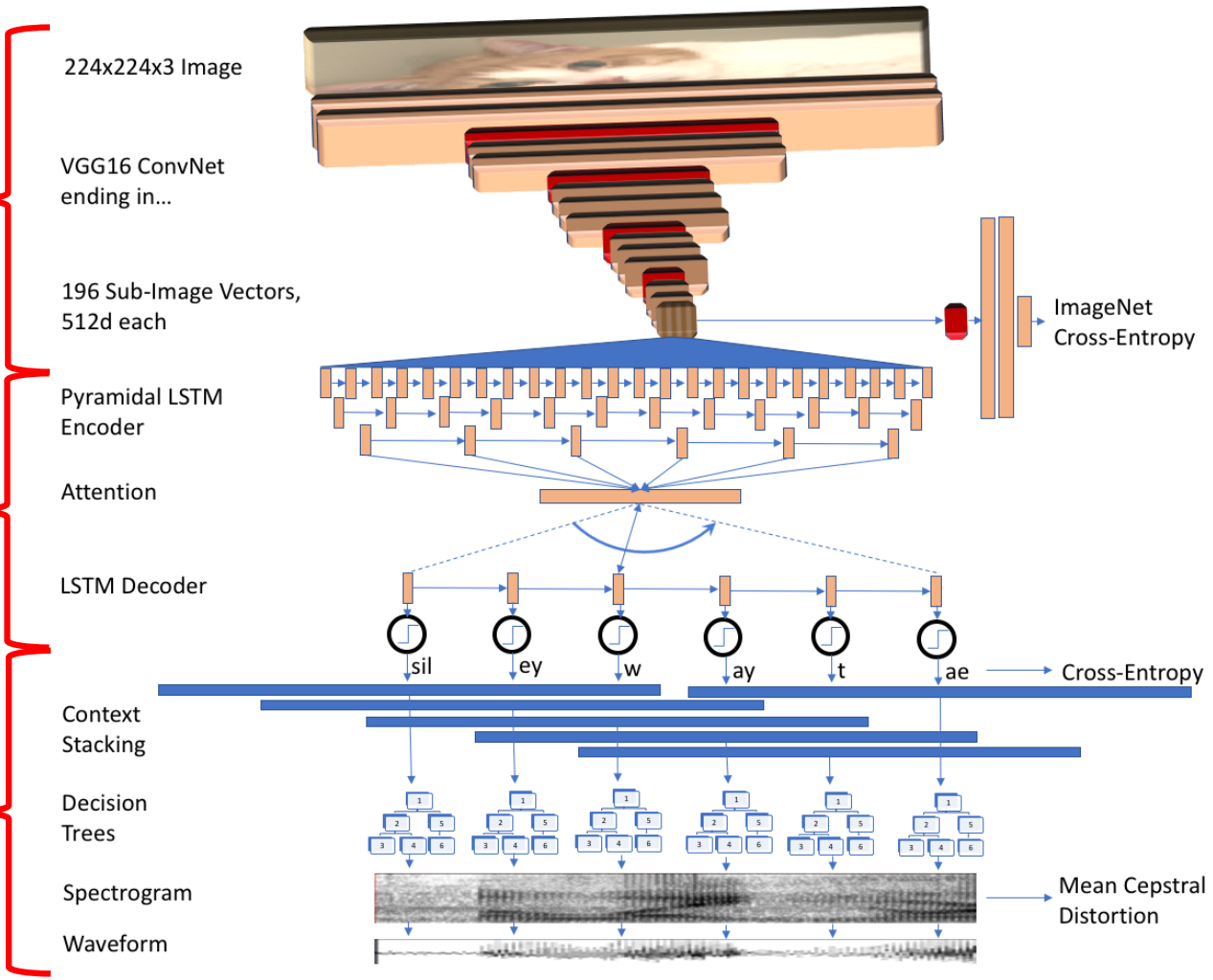


image2speech system components

- Image Representation:
  - Very large scale convolutional neural net VGG16, invented by Simonyan & Zisserman, I used the code published by Frossard (VGG16)
- Machine Translation:
  - Convert image vectors into phone-like speech segments using a sequence-to-sequence neural machine translator with attention, implemented using XNMT (Neubig et al.)
- Speech Synthesis:
  - Audio training data are segmented using Kaldi with whatever phone-like units are available
  - ClusterGen learns to map phone symbols to MFCC frames using a random forest (Black)

Image Representation: CNNFEAT $\vec{s}_{mn}$

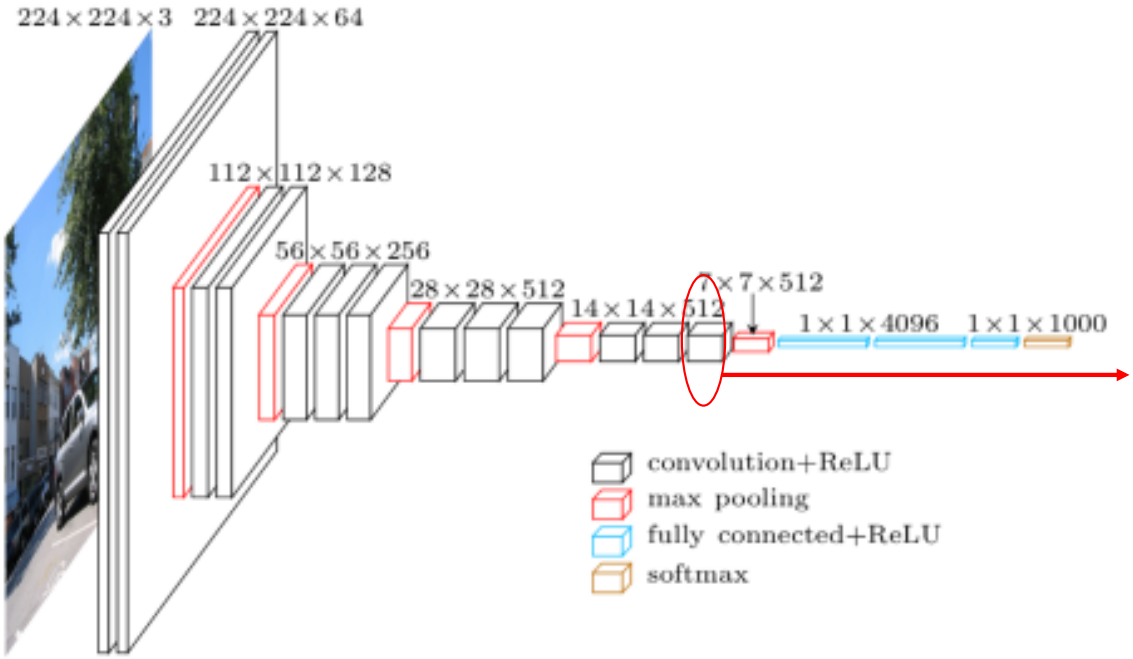
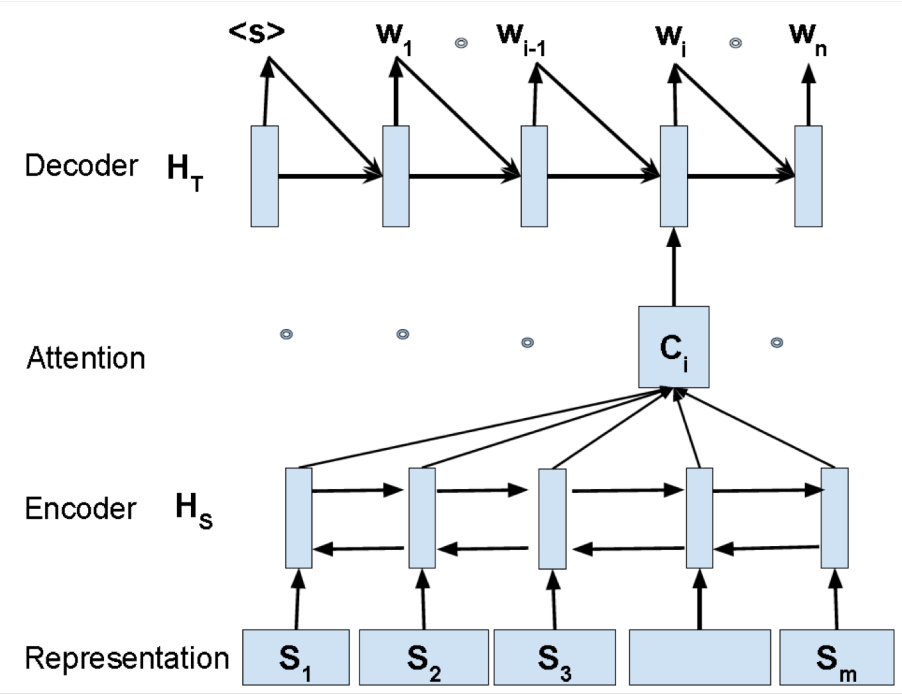


Figure copied without permission from Simonyan & Zisserman, 2014.

- ImageNet = >500 images/noun of each of the nouns in WordNet.
- VGG = 13-layer CNN + 2-layer FCN, trained on 14m images, covering the 1000 most numerous nouns, 92.7% top-5 test accuracy.
- **CNNFEAT: 196 feature vectors/image, 512d/vector, from the last CNN layer. Each receptive field covers about 40x40 pixels in the original 224x224 image.**

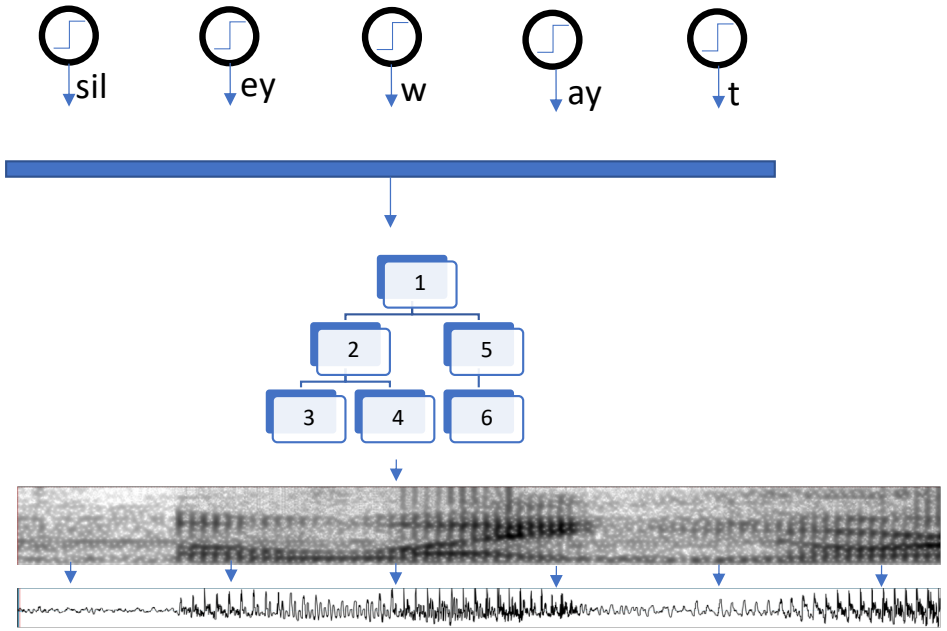
Machine Translation: images to phones



- Input Representation: 196 vectors/image
- Encoder: PyramidalLSTM with one 128d state vector. Sequence is row-wise raster scan of the image.
- Attention: StandardAttender, 128d input, 128d state vector, N hidden nodes
- Decoder: MlpSoftmaxDecoder, 3 layers, 1024d hidden vectors
- Output Representation: synthetic phones (MSCOCO), force-aligned phones (flickr8k), or acoustic unit discoveries (both)

Figure copied without permission from Duong, Anastasopoulos, Chiang, Bird & Cohn, NAACL-HLT 2016.

Speech Synthesis: phones to speech



- Input Representation: phone + context, context includes +/- 2 phones
- Regression: random forest is trained to generate MFCC+d+dd vectors from phone-context input vectors, using force-aligned audio corpus.
- Output: spectrogram is converted to audio using Griffin-Lim algorithm.

flickr8K: American phones





- Reference 1: “The boy +um+ laying face down on a skateboard is being pushed along the ground by +laugh+ another boy.”
- Reference 2: “Two girls +um+ play on a skateboard +breath+ in a court +laugh+ yard.”
- Hypothesis (128d attender): SIL +BREATH+ SIL T UW M EH N AA R R AY D IX NG AX R EH D AE N W AY T SIL R EY S SIL
- Hypothesis (64d attender): SIL +BREATH+ SIL T UW W IH M AX N W AO K IX NG AA N AX S T R IY T SIL
- Reference 1: “A boy +laugh+ in a blue top +laugh+ is jumping off some rocks in the woods.”
- Reference 2: “A boy +um+ jumps off a tan rock.”
- Hypothesis (128d attender): SIL +BREATH+ SIL EY M AE N IH Z JH AH M P IX NG IH N DH AX F AO R EH S T SIL
- Hypothesis (64d attender): SIL +BREATH+ SIL EY Y AH NG B OY W EY R IX NG AX B L UW SH ER T SIL IH Z R AY D IX NG AX HH IH L SIL

Images and Reference Texts: Hodosh, Young & Hockenmaier, 2013.
Natural Waveforms: Harwath and Glass, 2015

MSCOCO: British phones





Images and Reference Texts: MSCOCO

- Reference 1: “A group of men enjoying the beach, standing in the waves or surfing.”
- Reference 2: “A group of people standing on a beach next to the ocean.”
- Hypothesis (64d attender): # @ g r uu p @ v p ii p l= s t a n d i ng o n @ b ii ch # #

- Reference 1: “A, a black and white photo of a fire hydrant near a building.”
- Reference 2: “Aa, a fire hydrant that is out next to a house.”
- Hypothesis (64d attender): # @ p @@ s n= w oo k i ng @ t @ m e d^ l= d au n @ n d @ r e d f ai r h ai d r @ n t # #

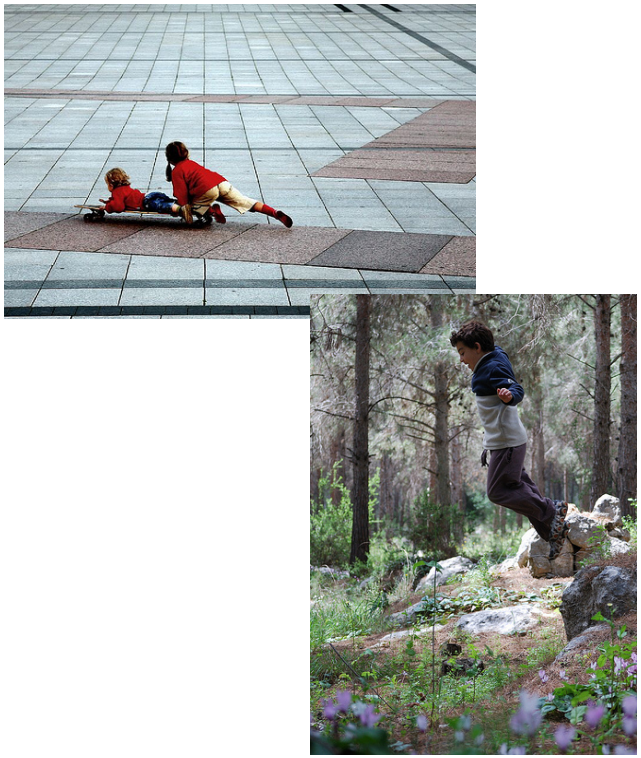
Numbers

Dataset	Phone Set	Dev BLEU (↑ better)	Dev UER (↓ better)	Test BLEU (↑ better)	Test PER (↓ better)
Flickr8k	Words	4.7	91.3	3.7	130
Flickr8k	American (L1) phones	**13.7**	**87.9**	**13.7**	**84.9***
Flickr8k	Dutch (L2) phones	5.4	115	6.1	101
MSCOCO 8k	Words	4.8		5.5	88.5
MSCOCO 8k	British (L1) phones	**15.1**		**16.3**	**78.8***
MSCOCO 8k	Unsupervised phones	2.2		1.4	123

BLEU = N-gram matching score of the generated phone string
PER = phone error rate of the generated phone string

* = better than chance

flickr8K:
Dutch phones



- Human Reference: “The boy laying face down on a skateboard is being pushed along the ground by another boy.”
- ASR Reference, Dutch phones: D EH M OW Y L AA AX N S EH L AX AX N S T EY K AX HH EY R AX N S W EY NG K AX Z AO L D EH R AA AA N W R AX N HH EH R AX N OW YD EH M OW Y
- Image2speech Hypothesis, Dutch phones: AA N M EY N AX N S L AA N S AA N AX N S AA L AX N S AA N AX N S AA L AX N S AA L AX N

- Human Reference: “A boy in a blue top is jumping off some rocks in the woods.”
- ASR Reference, Dutch phones: AX Z L AX N AA N AX N S K EY R M AO R K K L AA S S R W AA R
- Image2speech Hypothesis, Dutch phones: AA N M EY N IH N AX N W AA R AX N S T EH N IH NG AX N S T EH R AX N S T EH N S T EH N S T EH N S T EH N S T EH N S T EH N S T EH N S T EH N S T EH N S T EH N S T EH N S T EH N S T EH N D AX N

MSCOCO: unsupervised phones



Images and Reference Texts: MSCOCO

- Human Reference, Words: “A large balloon shaped to look like a giant vase with a flower.”
- ASR Reference, Unsupervised phones: a83 a6 a79 a68 a20 a66 a17 a42 a13 a0 a68 a62 a90 a69 a90 a55 a70 a52 a22 a79 a41 a2 a99 a26 a9 a97 a62 a41 a8 a0 a29 a39 a33 a50 a58 a99 a85 a60 a74 a78 a83 a37 a2 a55 a77 a20 a98 a38 a43 a98 a62 a34 a62 a70 a26 a97 a74 a20 a12 a51 a59 a56 a96 a6 a79 a65 a41 a66 a48 a87 a5 a65 a1 a9 a97 a31 a24 a14 a79 a68 a96 a19 a84 a14 a35 a19 a24 a49 a24 a89 a43 a7 a3 a4 a49 a24 a89 a5 a7 a71 a48 a87 a49 a39 a68 a1 a50 a89 a75 a72 a61 a17 a4 a31 a7 a79 a24 a35 a73 a91
- Image2speech Hypothesis, Unsupervised phones: a74 a23 a97 a27 a51 a59 a56 a37 a89 a74 a48 a59 a56 a37 a0 a97 a12 a62 a70 a26 a90 a64 a33 a26 a12 a62 a70 a26 a54 a83 a90 a69 a32 a22 a6 a90 a58 a77 a29 a78 a84 a26 a20 a66 a50 a99 a51 a59 a56

Speech Technology for Unwritten Languages

- speech-to-X and X-to-speech: apps with no writing
- speech-to-chat: generate a transcription in a language with no standard orthography, using chat alphabet
- Non-native (mismatched) transcription: makes sense if all humans have the same organs of production and perception
- Infer phones & tones in the unwritten language using two annotation languages
- image2speech = img2txt + text-to-speech - text
  - Works well (not perfectly) if the speech data can be force-aligned using a same-language ASR. Using a different-language ASR (e.g., Dutch) or a completely unsupervised ASR (acoustic unit discovery) does not (yet?) work well.