Aalto University
School of Electrical Engineering

Blizzard 2014 spoke task submission: Dual acoustic models and probabilistic cross-lingual speaker adaptation

R. Karhila[1], D. Gowda[1], M. Gibson[2], O. Watts[3], A. Suni[4], M. Kurimo[1]

[1] Aalto University, [2] Nuance Communications inc. [3] University of Edinburgh, [4] University of Helsinki

June 4th 2014

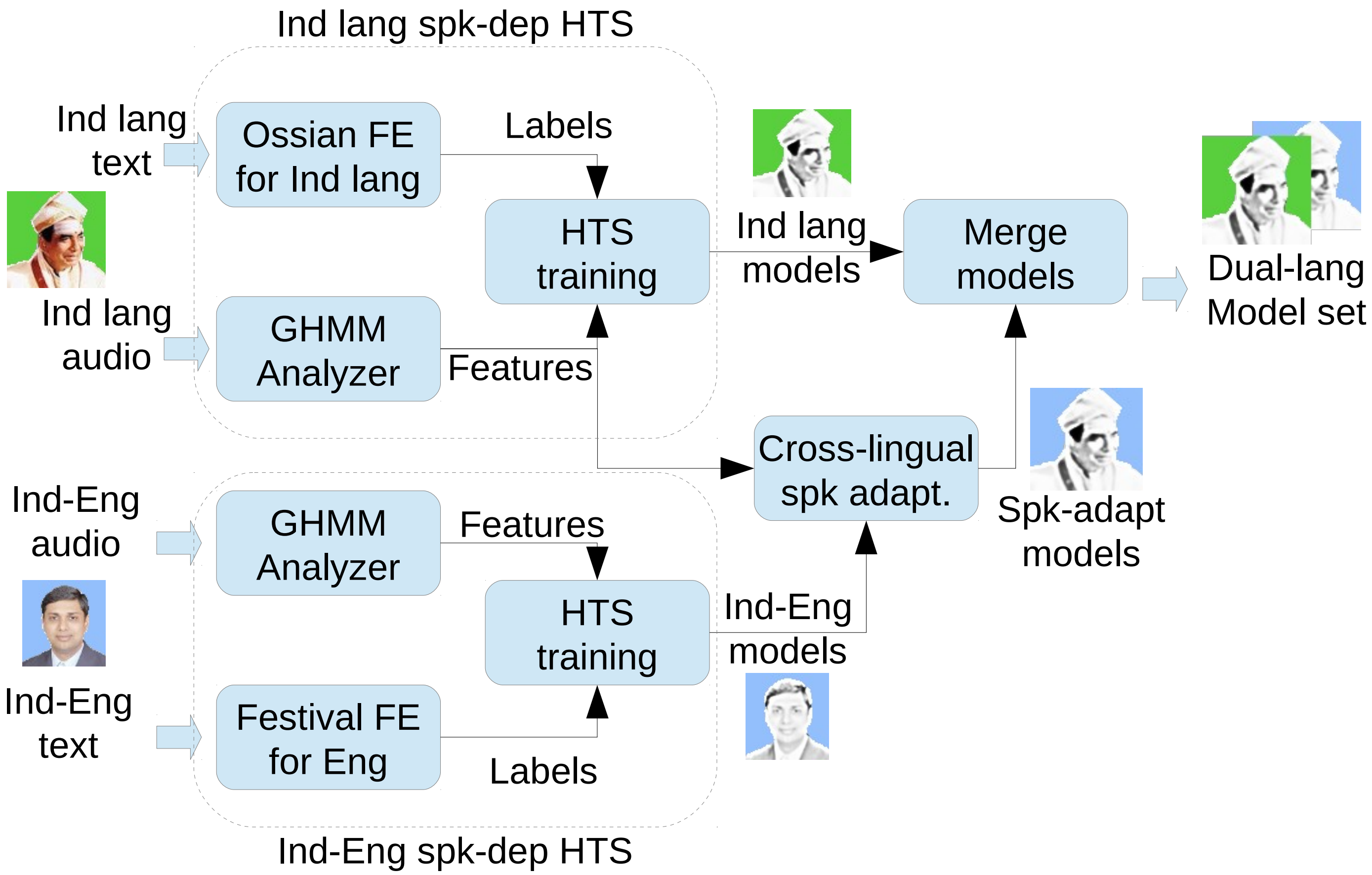
Outline

- Blizzard 2014 Spoke Task - Overview
- Dual language synthesis – overview
- Label generation
  - Method 1: Transcribe English in Indian script
  - Method 2: Dual front-end with filler words
- Acoustic model training
  - Speaker-dependent models
  - Cross-lingual adaptation
- Model set merging and synthesis
- Spoke Task - Demo

Blizzard 2014: Spoke Task

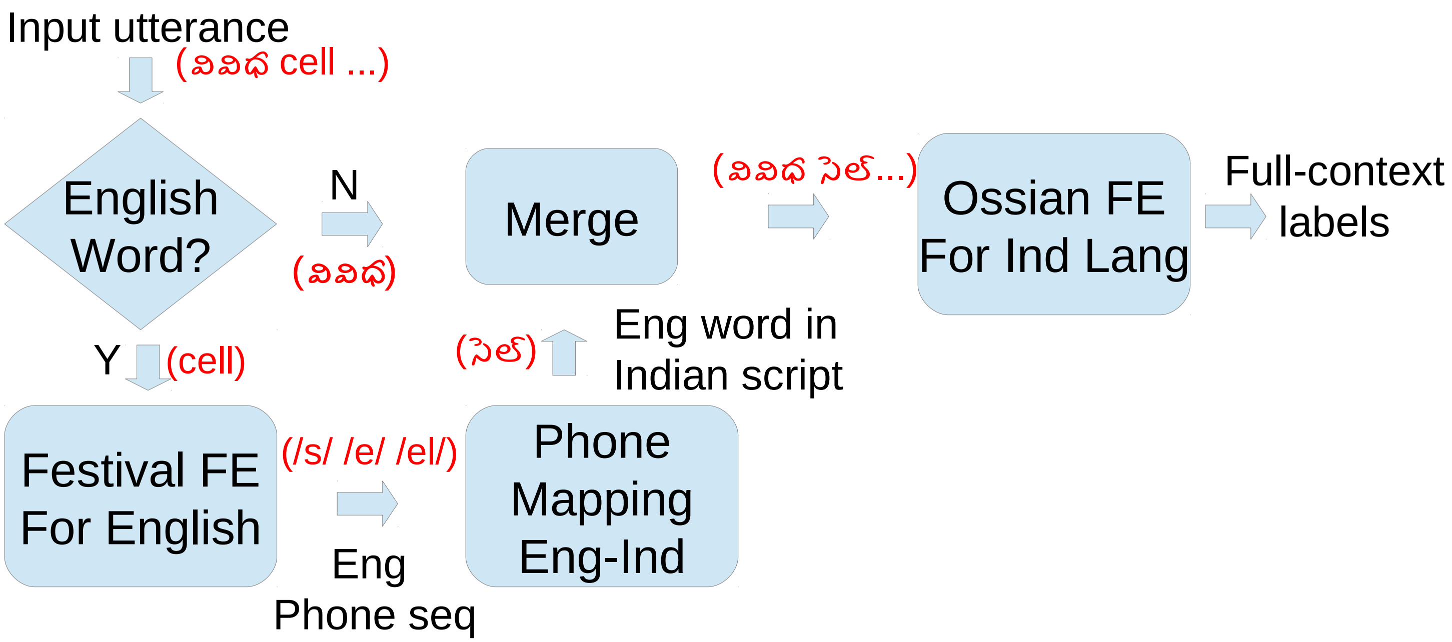
- Task: To sythesize dual-language utterances, primarily a native language (Indian) intersperced with words from a non-native language (English)
  - Training data
    - – Single speaker data only in Indian language (a few hundred utterances)
      - Example: "प्रसिद्घ कबीर अध्येता, पुरुषोत्तम अग्रवाल का यह शोध आलेख, उस रामानंद की खोज करता है “
    - – Audio data (16kHz, 16 bits) along with text in Indian script (UTF-8)
  - Test data
      - Example: “Under 19 cricket world cup में सोमवार को अफ़गानिस्तान ने ऑस्ट्रेलिया को हराकर, बड़ा उलटफेर किया है”

Dual-language synthesizer



Label generation - Method 1: Eng-to-Ind transcription

- Transcribe all English words in the target Indian langauge script and use the OSSIAN front-end (FE)



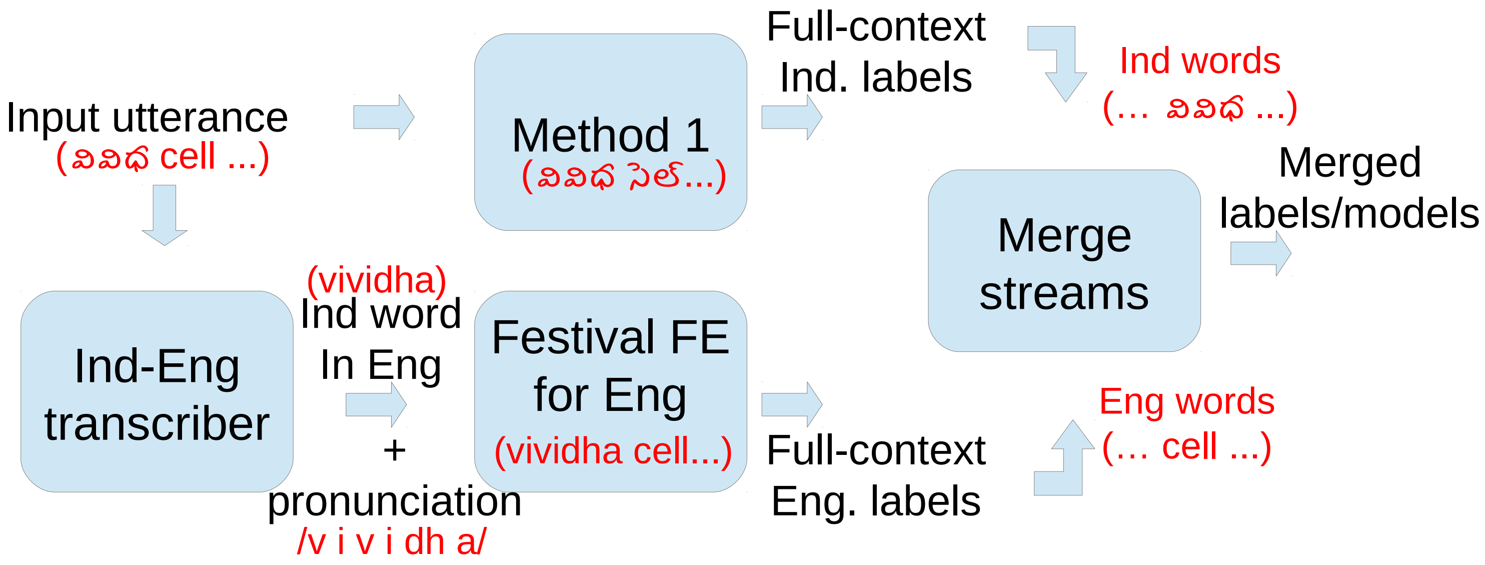
Aalto University
School of Electrical Engineering

Issues with Method 1: Eng-to-Ind transcription

- Works best for Telugu
  - Telugu script most phonetic
- Worst for Tamil
  - Context dependent phonemes in Tamil
  - Eg: single phoneme [k] represents both phones /k/ and /g/
- Not so good for Hin, Guj, Raj, and Asm.
  - Schwa insertion an isuue in Hin, Guj, and Raj.
  - Not able to assess Assamese

Label generation - Method 2: Dual front-end with filler words

- Use independent front-ends for English and the Ind. Lang., with filler words and merge the labels



Input utterance
(వివిధ cell ...)

Method 1
(వివిధ సెల్...)

Full-context
Ind. labels

Ind words
(… వివిధ ...)

Merge
streams

Merged
labels/models

Ind-Eng
transcriber

(vividha)
Ind word
In Eng
+
pronunciation
/v i v i dh a/

Festival FE
for Eng
(vividha cell...)

Full-context
Eng. labels

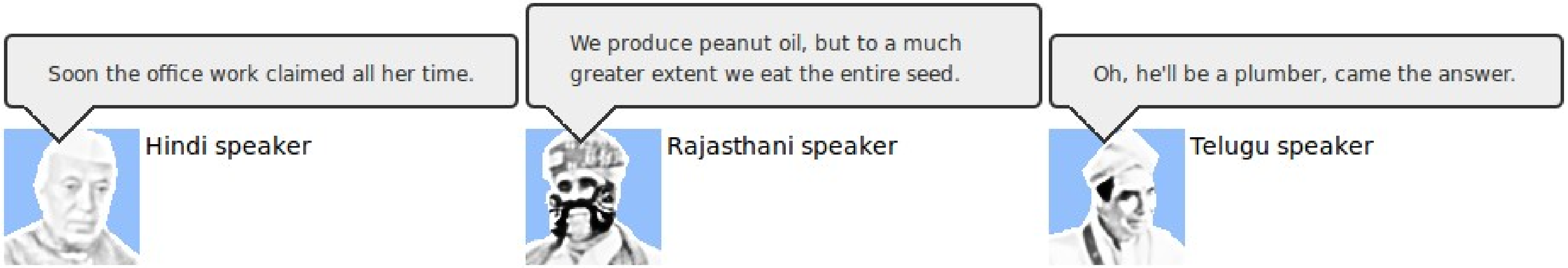
Eng words
(… cell ...)

Acoustic model training 1: Indian language speaker dependent HTS

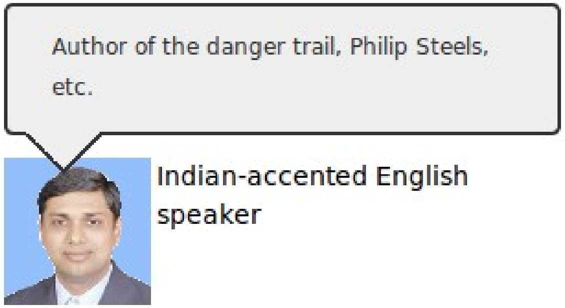
- Speaker dependent models for 3 different Indian languages – Hindi, Rajasthani and Telugu
- Training data (audio + UTF-8 text)
  - No. of uterrances - Hindi: 875, Rajasthani: 1369, Telugu: 1470
- GlottHMM acoustic modeling with Ossian front-end
- Features – LSFs:30, LSFsource:10, HNR:5, Gain:1, F0:1
- Trained with modified UEDIN Blizzard 2010 scripts.
  - Number of reclusterings = 3

Acoustic model training 2: Cross-lingual adaptation 0/4

- We want to create an English model set for each of the Indian language speakers:

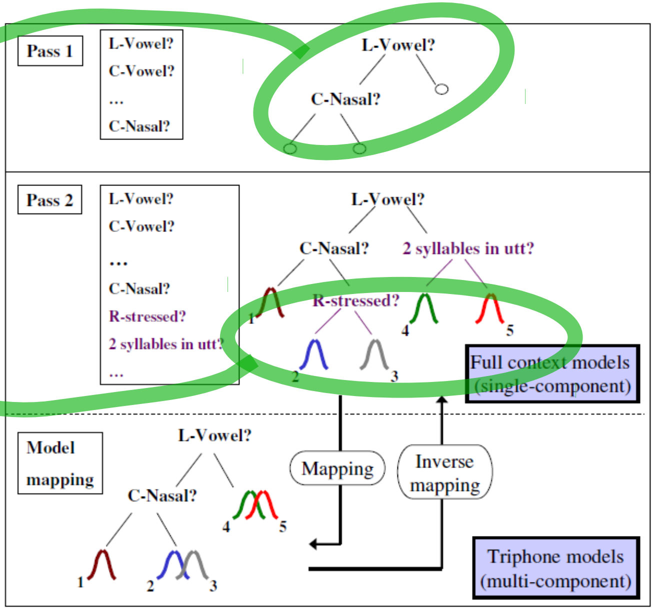

Soon the office work claimed all her time.
Hindi speaker
We produce peanut oil, but to a much greater extent we eat the entire seed.
Rajasthani speaker
Oh, he'll be a plumber, came the answer.
Telugu speaker

- We'll use the Arctic KSP speaker's Indian accented English data as a starting point


Author of the danger trail, Philip Steels, etc.
Indian-accented English speaker

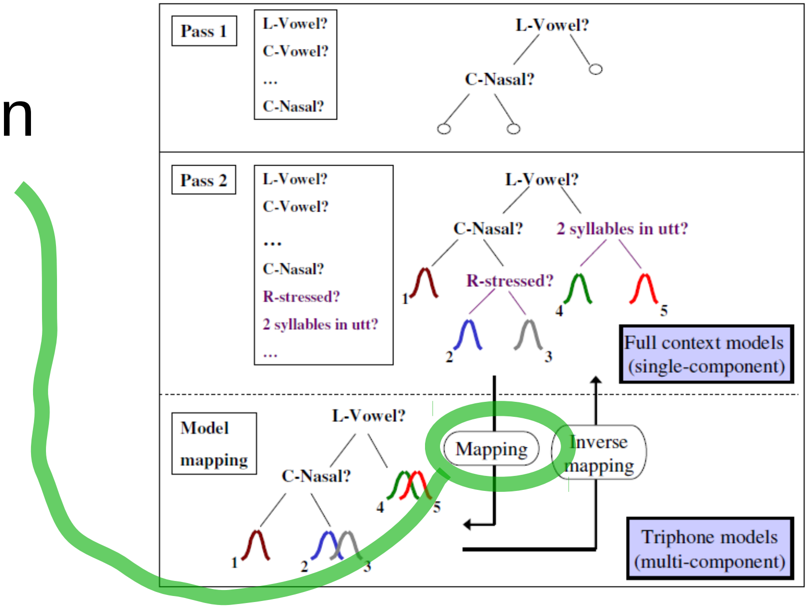
Acoustic model training 2: (Ind-Eng HTS) Cross-lingual adaptation 1/4

- English speaker-dependent model trained with Matt Gibson's code for two-pass decision tree generation:
- Top part of tree is populated by ASR-style questions related to triphone contexts – No leaf nodes!
- Bottom part of tree contains TTS style questions about quinphone context, stress, position in phrase etc... and leaf nodes with Gaussians



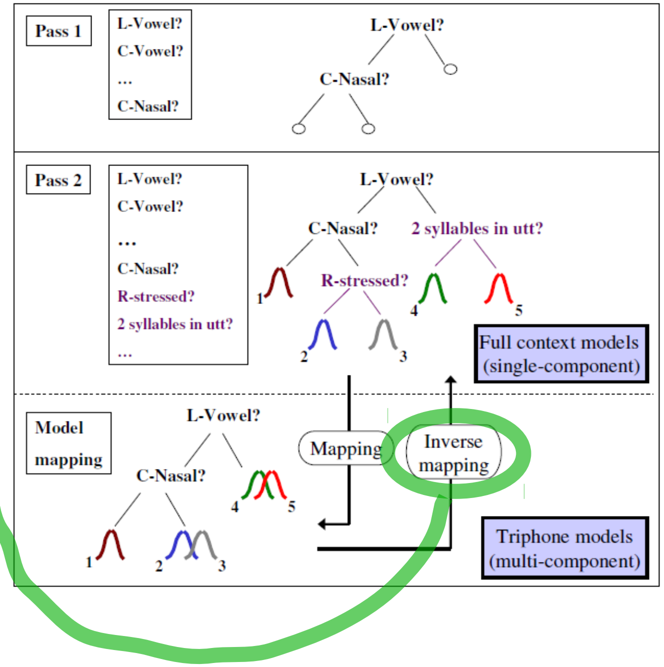
Acoustic model training 2: Cross-lingual adaptation 2/4

- After training, the Gaussians of the TTS model are collected based on the triphone contexts to make an ASR style model set
- This model set can be used to decode audio



Acoustic model training 2: Cross-lingual adaptation 3/4

- Indian language speakers' utterances are decoded using the triphone set
- A simple phoneme loop is used as the language model
- Triphone labels are then aligned using the different associated full-context TTS-models as alternative pronunciations for each triphone.
- This gives a mapping between Indian language audio and English TTS-models and allows adaptation to be carried out normally.



Acoustic model training 2: Cross-lingual adaptation 4/4

- Potential trouble
  - Lack of data for an average voice
  - We're adapting a speaker-dependent voice
- But at least we have a fair amount of training data for adaptation
- Could complete only 3 out of 6 Indian languages
  - Hindi, Rajasthani and Telugu completed. Stay tuned for samples!
  - Tamil: Lack of an Indian-English female voice
  - Gujarathi: Bad cross-lingual adaptation
  - Assamese: Bad speaker dep. voice for Indian lang.

Merging model sets

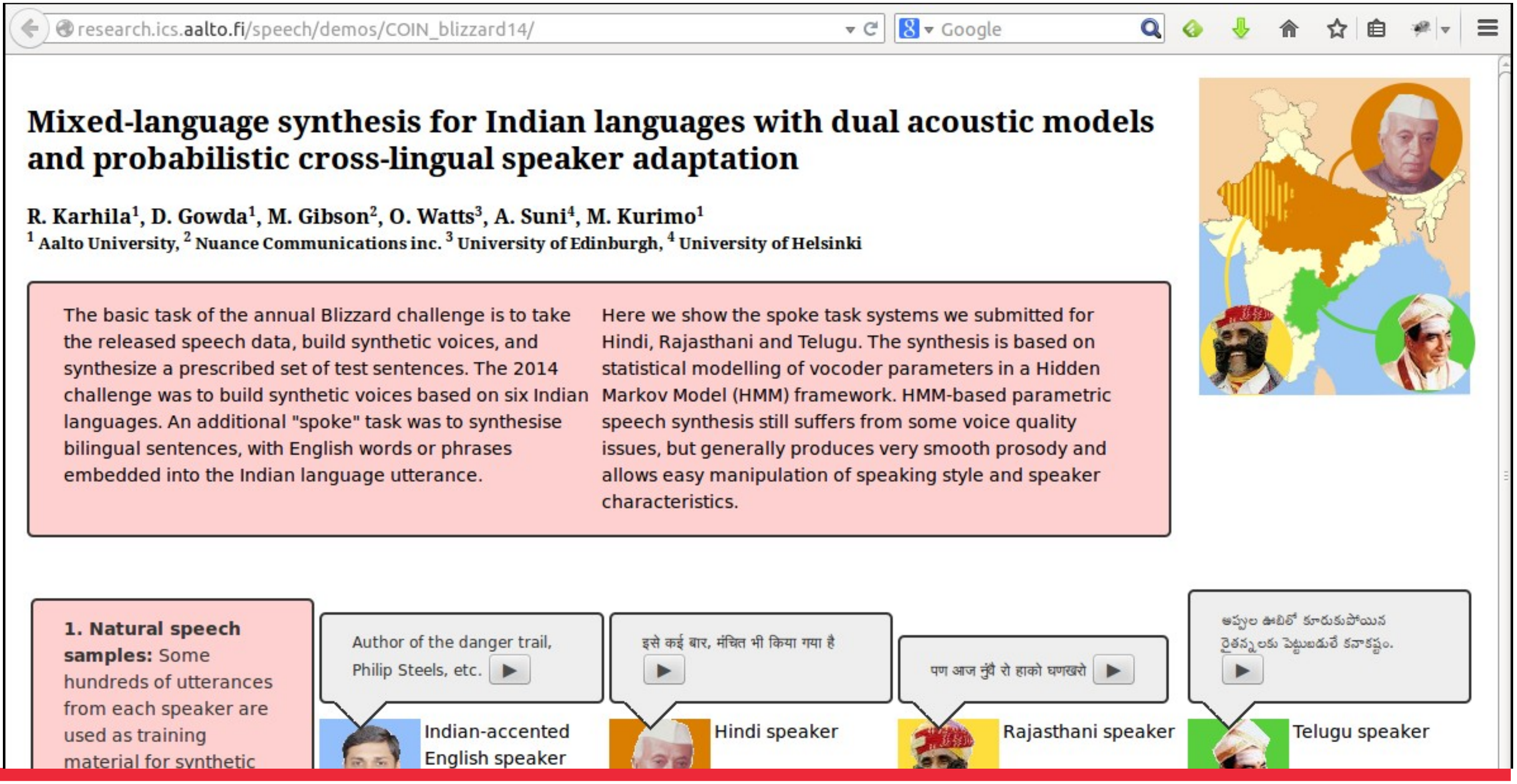
- Unseen models need to be synthesised using the decision trees, independently for both languages.
- The required models from both language model sets are concatenated:
  - Macros are renamed to retain uniqueness
  - Duplicate headers are removed
- Otherwise synthesis is done in a normal way

```
~o
<STREAMINFO> 6 93 30 15 1 1 1
<MSDINFO> 6 0 0 0 1 1 1
<VECSIZE> 141<NULLD><USER><DIAGC>
~t "TrP_10"
<TRANSP> 7
...
~p "hnr_s4_1440"
<STREAM> 3
<MEAN> 15
...
~h "-
s+/0:v/1:ee/2:s/3:u/4:k/5:15/6:11/7:35/8:25/9:23/10:61/11:56
/12:65/13:4/14:66/15:64/16:57/17:64/18:8/19:5/20:1/21:3/22:2
/23:2/24:2/25:2/26:5/27:0/28:0/29:15/30:0/31:10/32:0/33:11/3
4:3/35:15/36:7/37:10/38:0/39:0/40:1/41:0/42:5/43:0/44:0/45:3
6/46:0/47:26/48:0/49:4/50:0/51:29/52:4/53:0/54:8/55:0/56:0/5
7:38/58:5/59:1/"
<BEGINHMM>
<NUMSTATES> 7
...
<ENDHMM>
~t "enTrP_10"
<TRANSP> 7
...
~p "en(flow_s2_301)169"
<STREAM> 2
<MEAN> 30
...
~h "e~n-t^+@=n:4_1/A/0_0_2/B/1-1-4:3-2&4-8#1-2$2-2>0-2<2-7|
e/C/0+0+3/D/content_1/E/content+4:2+6&2+5#1+1/F/content_2/G/
0_0/H/11=7:1=1&L-L%/I/0_0/J/11+7-1"
<BEGINHMM>
...
```

Aalto University
School of Electrical
Engineering

Demo page

http://research.ics.aalto.fi/speech/demos/COIN_blizzard14/



Mixed-language synthesis for Indian languages with dual acoustic models and probabilistic cross-lingual speaker adaptation

R. Karhila[1], D. Gowda[1], M. Gibson[2], O. Watts[3], A. Suni[4], M. Kurimo[1]
[1] Aalto University, [2] Nuance Communications inc. [3] University of Edinburgh, [4] University of Helsinki

The basic task of the annual Blizzard challenge is to take the released speech data, build synthetic voices, and synthesize a prescribed set of test sentences. The 2014 challenge was to build synthetic voices based on six Indian languages. An additional "spoke" task was to synthesise bilingual sentences, with English words or phrases embedded into the Indian language utterance.

Here we show the spoke task systems we submitted for Hindi, Rajasthani and Telugu. The synthesis is based on statistical modelling of vocoder parameters in a Hidden Markov Model (HMM) framework. HMM-based parametric speech synthesis still suffers from some voice quality issues, but generally produces very smooth prosody and allows easy manipulation of speaking style and speaker characteristics.

1. Natural speech samples: Some hundreds of utterances from each speaker are used as training material for synthetic

Author of the danger trail, Philip Steels, etc. ▶
Indian-accented English speaker

इसे कई बार, मंचित भी किया गया है ▶
Hindi speaker

पण आज नुंवै रो हाको घणखरो ▶
Rajasthani speaker

అప్పుల ఊబిలో కూరుకుపోయిన రైతన్నలకు పెట్టుబడులే కనాకష్టం. ▶
Telugu speaker

Aalto University
School of Electrical Engineering

Issues to Ponder Over?

- How can a simple thing like this work?
- Role of filler words in label generation
  - for smooth transition
  - word count, accurate phone/context at boundaries
  - what else?
- How significant is syllable and stress related information for Indian languages?

References

1. Gibson, M.; Byrne, W., "Unsupervised Intralingual and Cross-Lingual Speaker Adaptation for HMM-Based Speech Synthesis Using Two-Pass Decision Tree Construction," Audio, Speech, and Language Processing, IEEE Transactions on , vol.19, no.4, pp.895,904, May 2011
2. Gibson, M.; Hirsimäki, T.; Karhila, R.; Kurimo, M.; Byrne, W., "Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis using two-pass decision tree construction," Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on , vol., no., pp.4642,4645, 14-19 March 2010
3. Simple4All Ossian front-end, http://homepages.inf.ed.ac.uk/owatts/ossian/html/index.html
4. T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-Based Speech Synthesis Utilizing Glottal Inverse Filtering", in IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 1, pp. 153-165, 2011.
5. Tokuda, K.; Nankaku, Y.; Toda, T.; Zen, H.; Yamagishi, J.; Oura, K., "Speech Synthesis Based on Hidden Markov Models," Proceedings of the IEEE , vol.101, no.5, pp.1234,1252, May 2013
6. Heiga Zen, Keiichi Tokuda, Alan W. Black, Statistical parametric speech synthesis, Speech Communication, Volume 51, Issue 11, November 2009, Pages 1039-1064
7. Tokuda, K., Zen, H., Yamagishi, J., Black, A., Masuko, T., Sako, S., Toda, T., Nose, T., Oura, K., 2008. The HMM-based speech synthesis system (HTS). <http://hts.sp.nitech.ac.jp/>.