Corpora

OGI MLTS

- Telephone speech in 11 languages
  - English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil, Vietnamese
- ~2000 speakers, 38.5 hours total
- Fixed vocabulary plus an open vocabulary section
- (LDC 1994)

OGI MLTS Protocol

1. What is your native language?
2. What language do you speak most of the time?
3. What language do you speak at home?
4. How old are you?
5. What is your date of birth?
6. Are you male or female?
7. Were you born and raised in the United States?
8. What city and state did you spend most of your childhood?
9. What is your zipcode?
10. What area code are you calling from?
11. What day is today?
12. What time is it?
13. For each of the following descriptions, we will record the first ten seconds of your answer. Begin speaking at the beep. A second beep will indicate when we have finished recording your answer to each question.
(pause)
14. Describe the route you take to work or to the store.
15. Tell us something that you like about your hometown.
16. Tell us about the climate in your hometown.
17. Describe the room you are calling from.
18. Describe your most recent meal.
19. We now want you to talk for a longer period of time. We do not care what you say as long as you keep talking. You can tell us anything about yourself, your hobbies and interests, the city that you live in, and the sports that you like. Or you can make up a story, tell a fairy-tale or recite a poem. You will have 1 minute to speak. We will now give you 10 seconds to think about what to say. Please do not read anything, we would prefer you make something up.
(pause)
20. Please begin talking at the beep. You will hear a second beep when you have 10 seconds left.
21.You have 10 seconds to complete your story.
22. If you are calling from a touch tone phone, please push the number 2 button.

23. Would you like to receive a gift certificate for MacDonalds or for TCBY frozen yogurt?
24.Thank you for your participation. If you would like a gift certificate please leave your name, address, and gift certificate selection. Your name and address will be kept confidential.

Others

- European Parliament transcriptions
  - 5+ languages
- EUROM1
  - >60 speakers in 7 European languages
- Speecon
  - Dates, digit sequences, phonetically rich utterances
  - >550 adult speakers per language
  - 4 mics, lexicon included
  - 50k EUR per language

GlobalPhone

- Read newspaper speech in over 15 languages
  - 16kHz mono in quiet, but not studio, conditions
  - ~100 speakers and >20 hours per language
  - Romanised and original script word transcriptions
  - Language models and lexica “available”

	Speakers	hours	words
Arabic	170	35	i.p.
Ch-Mandarin	132	31	263k
Ch-Shanghai	41	10	95k
Croatian	92	16	120k
Czech	102	29	220k
French	94	25	250k
German	77	18	151k
Japanese	144	34	268k
Korean	100	21	117k
Portuguese	101	26	208k
Russian	106	22	170k
Spanish	100	22	172k
Swedish	98	22	184k
Tamil	49	i.p.	i.p.
Turkish	100	17	113k
Total	**1506**	**328**	**2331k**

	family
Arabic	Afro-Asiatic -> Semitic -> Central -> South -> Arabic
Ch-Mandarin	Sino-Tibetan -> Chinese
Ch-Shanghai	Sino-Tibetan -> Chinese
Croatian	Indo-European -> Slavic -> South -> Western
Czech	Indo-European -> Slavic -> West -> Czech-Slovak
French	Indo-European -> Italic -> Romance -> Italo-Western -> Western -> Gallo-Iberian -> Gallo-Romance -> Gallo-Rhaetian -> Oïl
German	Indo-European -> Germanic -> West -> High German -> German -> Middle German -> East Middle German
Japanese	Japanese
Korean	Language Isolate
Portuguese	Indo-European -> Italic -> Romance -> Italo-Western -> Western -> Gallo-Iberian -> Ibero-Romance -> West Iberian -> Portug
Russian	Indo-European -> Slavic -> East
Spanish	Indo-European -> Italic -> Romance -> Italo-Western -> Western -> Gallo-Iberian -> Ibero-Romance -> West Iberian -> Castili
Swedish	Indo-European -> Germanic -> North -> East Scandinavian -> Danish-Swedish -> Swedish
Tamil	Dravidian -> Southern -> Tamil-Kannada -> Tamil-Kodagu -> Tamil-Malayalam -> Tamil
Turkish	Altaic -> Turkic -> Southern -> Turkish

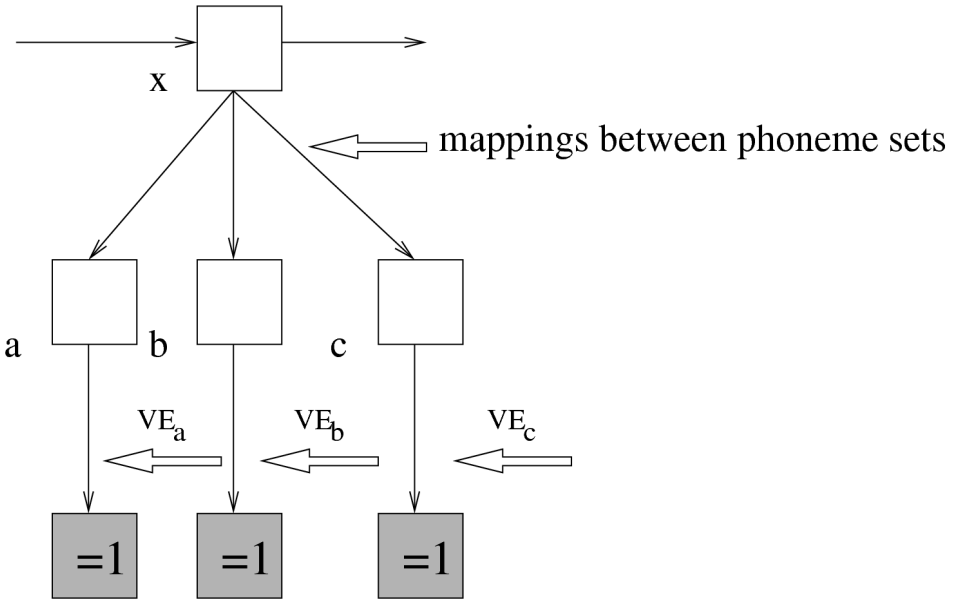
Language choice

- Mandarin
  - tone
- Arabic
  - pharyngeals
- German -> plus Swedish
  - consonant clusters
- Portuguese -> plus Spanish, French
  - nasals
- Russian -> plus Croatian, Czech
  - palatized sounds

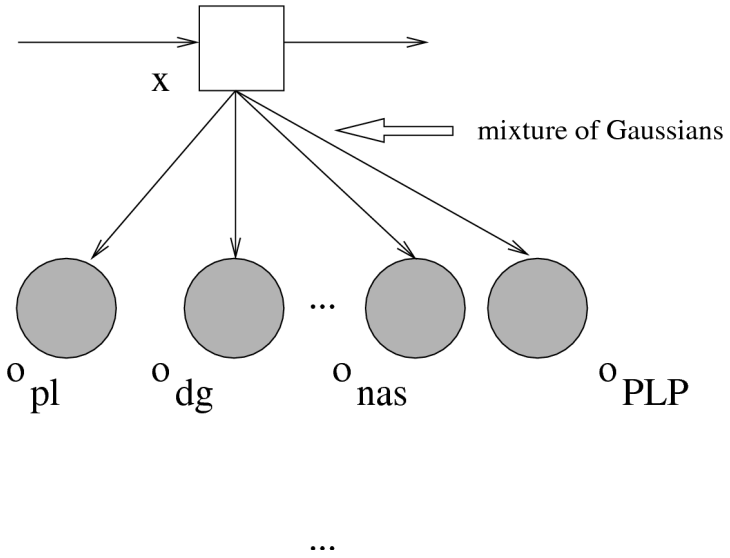
Plans

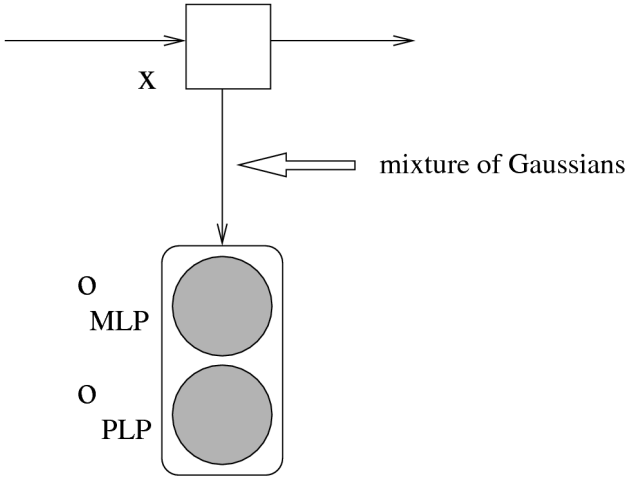
- Models
  - Hybrid ANN/HMM
  - Tandem
- Training scenarios
  - A → X
  - {A, B, C} → X
  - {A, B, X} → X

Hybrid



Tandem


x
mixture of Gaussians
o pl
o dg
...
o nas
o PLP
...


x
mixture of Gaussians
o MLP
o PLP

Plans...

- (Hybrid)
  - Learnt mapping from A-phones to X-phones
  - Learnt mapping from AFs to phones
- Tandem
  - Different factorings
  - Separate tying

Plans...

- Multi-task learning
  - MLPs could have both sub-word unit targets and language targets
    - N output units for phones/graphemes etc
    - M output units for language
    - N+M rather than N*M?