

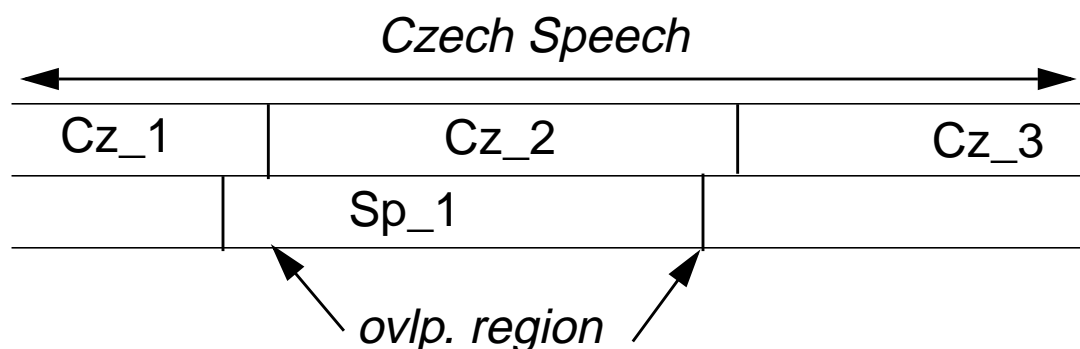
Mapping phones and states across languages through the confusion matrix

We assume we have a collection of well trained acoustic models in several source languages, we want to know:

- **How useful are source language models when recognizing a new language?**

The confusion matrix technique (applied to phones):

- Given a “small” collection of target data and its transcription, perform **automatic alignment**
- Given the set of acoustic models of a given source language do **phonetic recognition** on the same data
- Compare overlaps between the alignment (reference) and the outcome of the phonetic recognition (hypothesis): represent these overlaps in the form of a confusion matrix
- For each phone in the target language choose the best matching phone in the source language

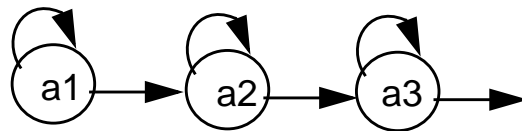


Extending phonetic mapping approach to state level

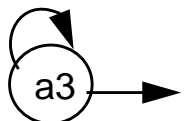
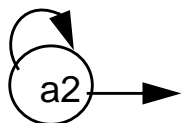
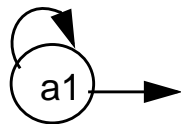
Some phones seem to be better candidates to be shared across languages than others

- Some phones seem to be composed of phonetic subunits that exist in other languages (e.g., diphthongs, ñ -> n_y)
- How can we find match these sub-units if the HMM are designed to represents monophones?

Model: "a"



will generate 3 models: "a1", "a2", "a3"



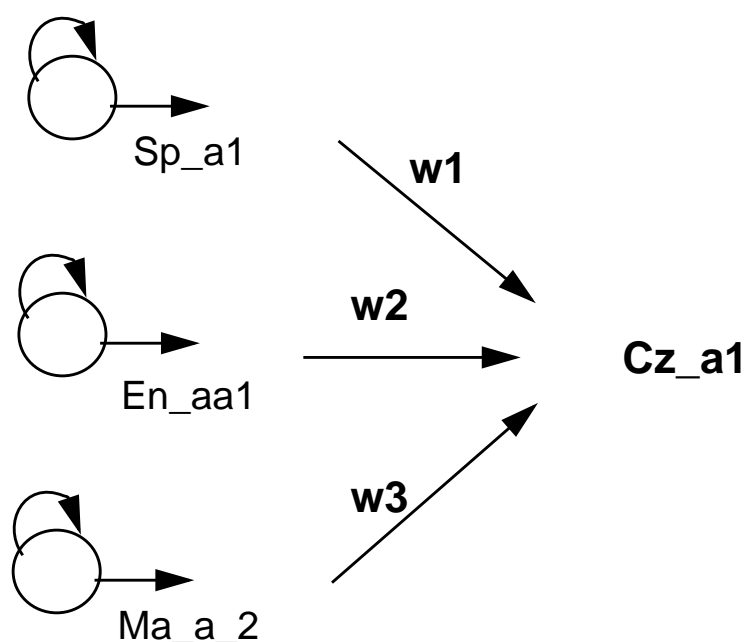
We can now treat states as monophones, and generate confusion matrices using the confusion matrix approach

Multiple state mapping and target state assembling

We now can find for every target language state/phone:

- Best matching state/phone from a given language
- Best state/phone from a set of languages
- N-best states/phones from a set of languages

From these mappings we can assemble target language states and HMMs by weighting the gaussian mixtures of the source states



We effectively assemble them in the following way:

$$p(o|Cza1) = \sum_i \sum_j w(i, j) c(j) N(o, m_{ij}, C_{ij})$$

Examples of state level mappings

a_2	MA_at_1	0.09	SP_a_1	0.08	EN_aw_1	0.05
a_3	SP_a_2	0.06	EN_ah_1	0.06	EN_eh_3	0.06
a_4	SP_a_3	0.09	EN_ah_3	0.07	MA_at_3	0.07
aa_2	EN_aw_1	0.18	EN_aa_1	0.16	EN_ah_1	0.08
aa_3	EN_aw_2	0.33	EN_aax_2	0.13	EN_aa_2	0.10
aa_4	EN_aax_3	0.25	EN_aw_3	0.17	MA_at_2	0.06
aw_3	MA_gt_1	0.33	EN_l_1	0.33	EN_ax_3	0.33
b_2	MA_b_1	0.17	EN_v_1	0.10	SP_b_1	0.09
b_3	SP_b_2	0.14	MA_d_1	0.12	MA_b_1	0.12
b_4	EN_b_3	0.12	EN_b_2	0.12	MA_d_2	0.09
c_2	MA_c_1	0.29	MA_q_1	0.12	MA_z_1	0.07
c_3	MA_c_2	0.40	EN_t_2	0.09	MA_z_2	0.07
c_4	SP_x_2	0.09	SP_x_3	0.08	MA_s_3	0.08
dj_2	SP_y_1	0.10	MA_U_3	0.08	EN_jh_1	0.08
dj_3	SP_y_1	0.22	SP_II_1	0.12	EN_d_2	0.12
dj_4	MA_j_3	0.21	SP_II_2	0.14	EN_d_3	0.09

Some Preliminary Results

Source(s)	level	method	n-best	%Accu.
E	Phone	manual	1	<20%
S	Phone	manual	1	<20%
E	Phone	matrix	1	31.69%
S	Phone	matrix	1	31.33%
E	State	matrix	1	35.25%
S	State	matrix	1	29.97%
M	State	matrix	1	20.31%
E+S+M	State	matrix	1	37.72%
E+S+M	State	matrix	3	44.23%

Directions:

- Consider confidence
- Reduction of Gaussians, Phonetic context
- What is the effect of the amount of target language data?
- Cross language- multiple **domain** experiments