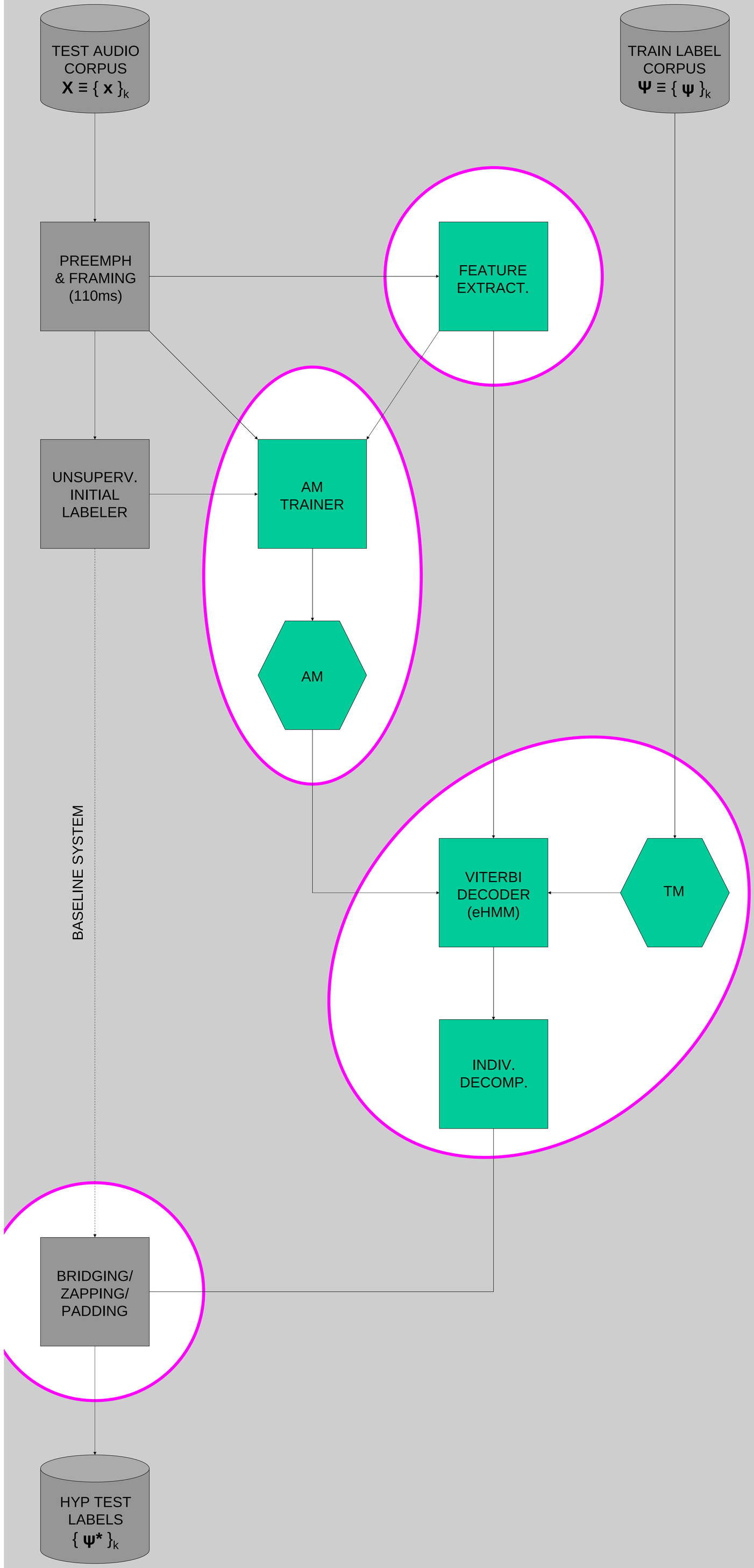
Implementation



Results

Baseline (ISL RT04s STT system)

	RT04s DEV			RT04s EVAL		
	MS	**FA**	**1 – F**	**MS**	**FA**	**1 – F**
Pre-smoothing	32.90	4.18	21.07	44.92	4.63	30.17
Smoothing	14.41	14.17	14.29	29.36	15.68	23.23

Feature Extraction

- 2 simple features **per participant**

	RT04s DEV			RT04s EVAL		
	MS	**FA**	**1 – F**	**MS**	**FA**	**1 – F**
ENE + ZCR	31.51	8.10	21.51	41.61	8.09	28.58

Acoustic Model Training

- 3 techniques for reusing or creating data mass for training
- Each vocal interaction state is modeled with **a one Gaussian**

	RT04s DEV			RT04s EVAL		
	MS	**FA**	**1 – F**	**MS**	**FA**	**1 – F**
Covariance Sharing	28.67	7.69	19.52	37.93	7.19	25.61
Channel Rotation	28.69	7.04	19.36	38.20	5.95	25.48
Overlap Synthesis	25.20	9.16	17.96	35.41	8.43	24.25

Vocal Interaction Transition Modeling

- The above assumes uniform transition probabilities
- A participant-independent model:

$$P(\Psi[t+1] \mid \Psi[t]) \approx P(\|\Psi[t+1]\|, \|\Psi[t+1] \times \Psi[t]\| \mid \|\Psi[t]\|)$$

The transition probability is a function of:

1. How many participants were speaking at time t;
2. How many participants were speaking at time t+1; and
3. How many of them are the same.

	RT04s DEV			RT04s EVAL		
	MS	**FA**	**1 – F**	**MS**	**FA**	**1 – F**
Viterbi Decoding	22.81	4.77	14.73	32.46	2.78	20.29

Bridging gaps, etc.

Performed to match the granularity of the ref segmentation

	RT04s DEV			RT04s EVAL		
	MS	**FA**	**1 – F**	**MS**	**FA**	**1 – F**
Smoothing	9.81	9.24	9.53	21.40	6.94	14.78