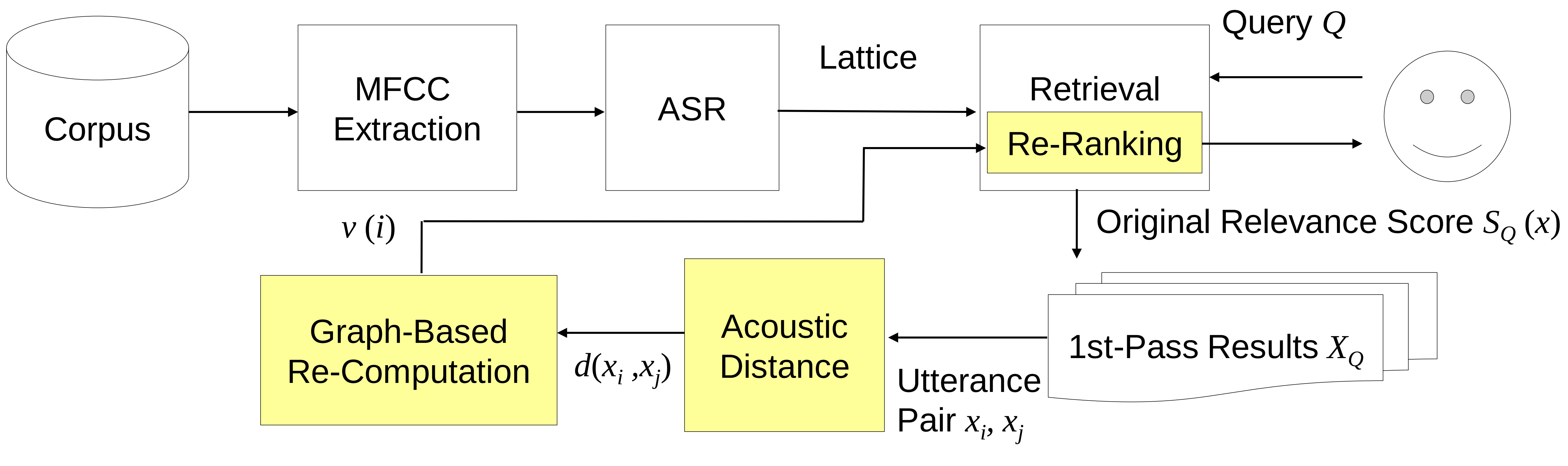
Improved Spoken Term Detection with Graph-Based Re-Ranking in Feature Space

Yun-Nung Chen, Chia-Ping Chen, Hung-Yi Lee , Chun-an Chan, and Lin-shan Lee

臺灣大學 National Taiwan University

Summary

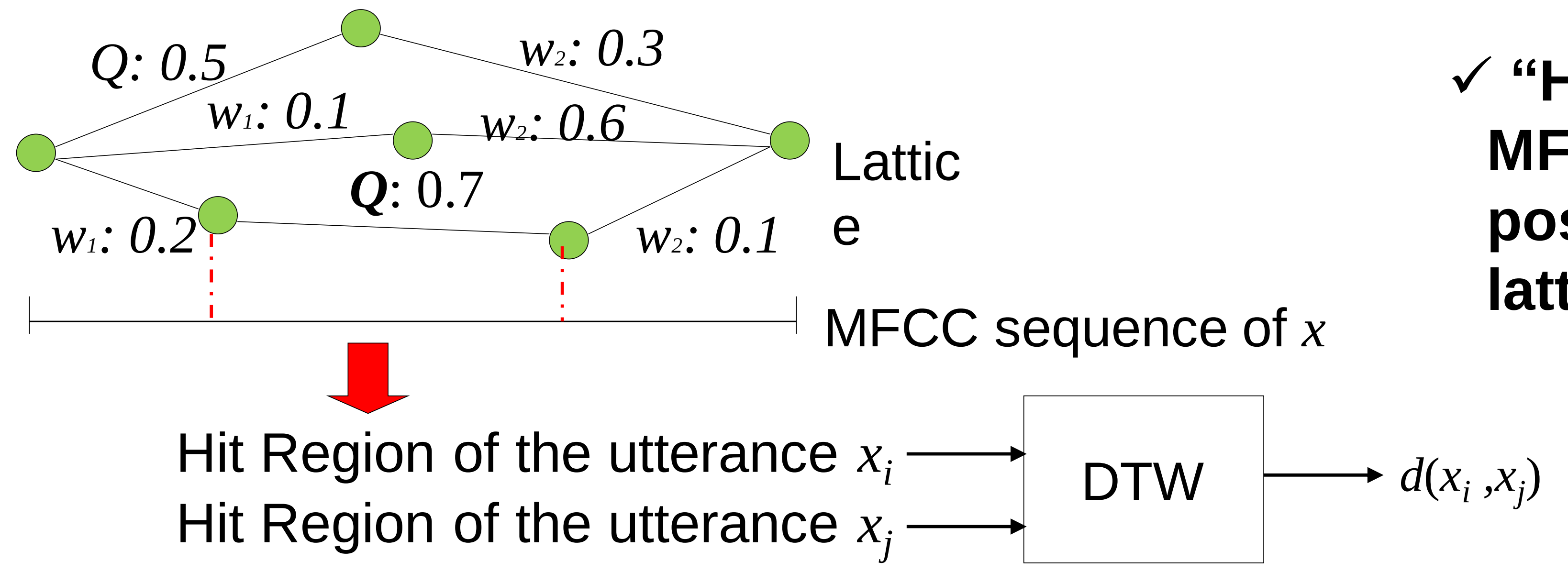
- **Using graph-based re-ranking to improve spoken term detection with acoustic similarity.**
- **With MLLR acoustic model, MAP improves from 55.54% to 67.38%. The relative improvement rate is 22.04%**



White blocks: original spoken term detection

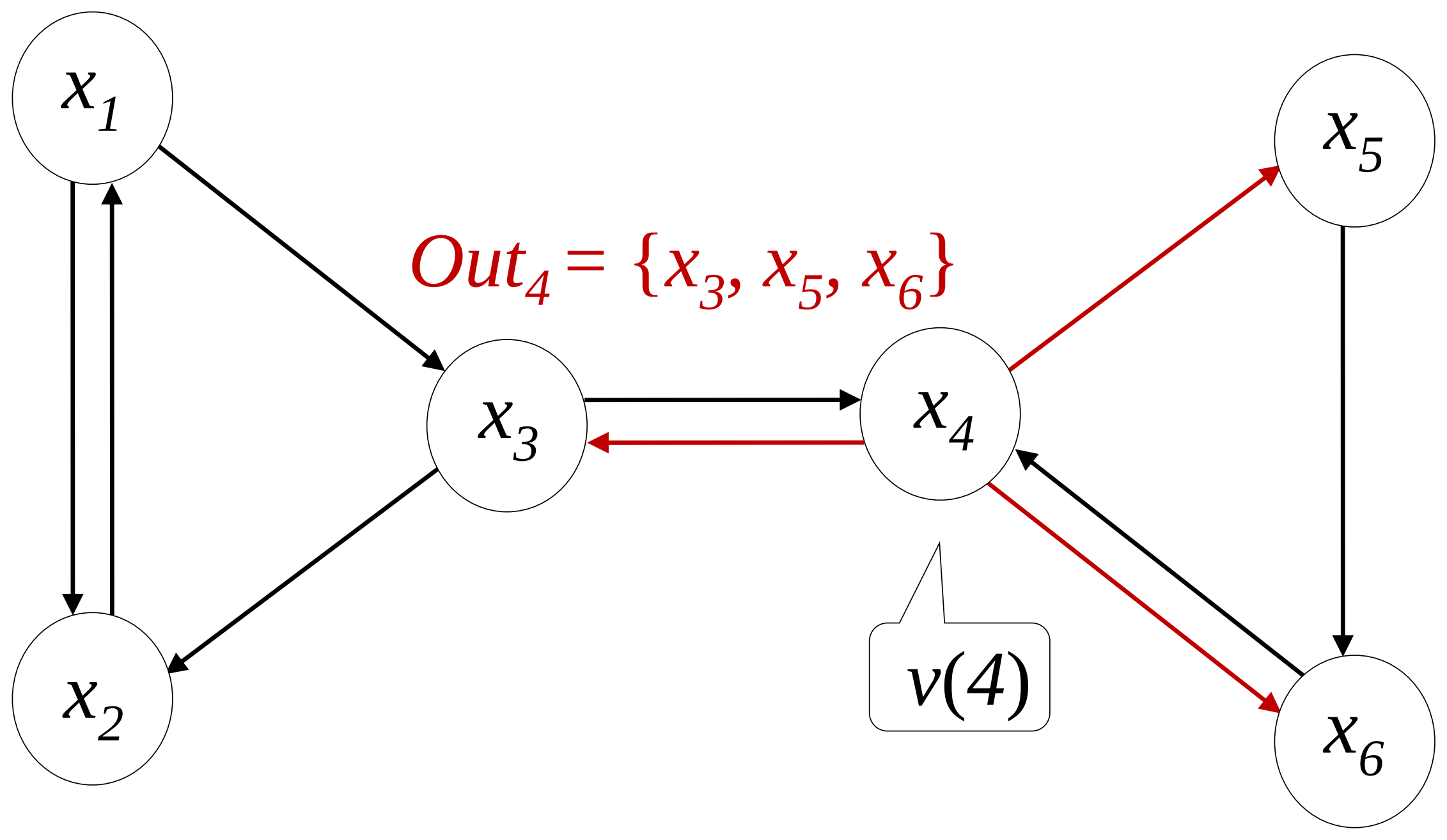
Acoustic Distance

- Compute acoustic distance $d(x_i, x_j)$ for each utterance pair x_i, x_j in first-pass result.



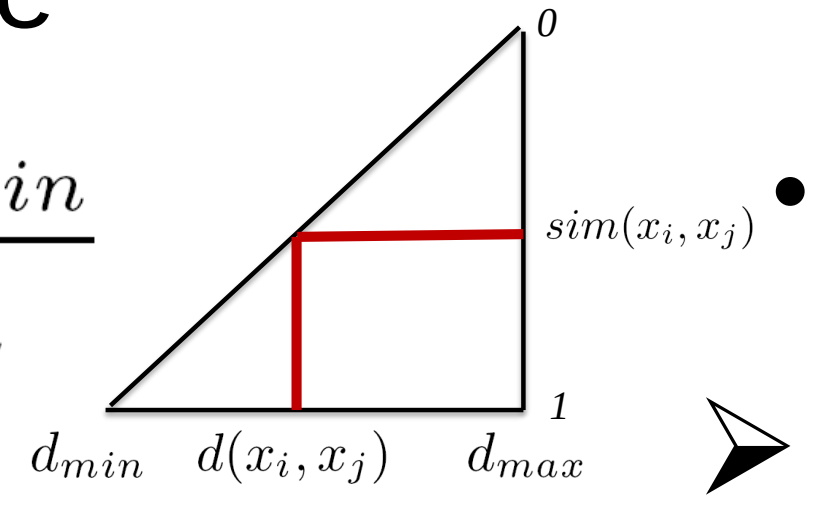
✓ **"Hit Region": the corresponding MFCC sequence which is most possible to be the query in the lattice.**

Graph-Based Re-Ranking with Acoustic Feature



- Node: first-pass retrieved utterance
- Edge: weighted by the similarity between the two utterances evaluated in feature space
- Compute similarity from distance

$$sim(x_i, x_j) = 1 - \frac{d(x_i, x_j) - d_{min}}{d_{max} - d_{min}}$$



- Normalized similarity

$$p(i,j) = \frac{sim(x_i, x_j)}{\sum_{x_k \in Out(v_i)} sim(x_i, x_k)}$$

➢ Modified Random Walk

$$v(i) = \frac{1}{\lambda}((1-\alpha)r(i) + \alpha \sum_{v_j \in Out_i} p(i,j)v(j))$$

$$r(i) = \frac{S_Q(x_i)}{\sum_{x_j \in X_Q} S_Q(x_j)}$$

- Normalized original relevance score

- Scores propagated from neighbors of node *i*

◆ ***v*(*i*) is higher when**

1) **Higher original relevance score**
2) **Similar to more utterances with higher scores**

$$\mathbf{v} = \frac{1}{\lambda}((1-\alpha)\mathbf{r} + \alpha \mathbf{P}^{\mathbf{T}}\mathbf{v}) = \frac{1}{\lambda}\mathbf{P}'\mathbf{v} \quad \text{(matrix form)}$$

- Solution: dominant eigenvector of *P'*

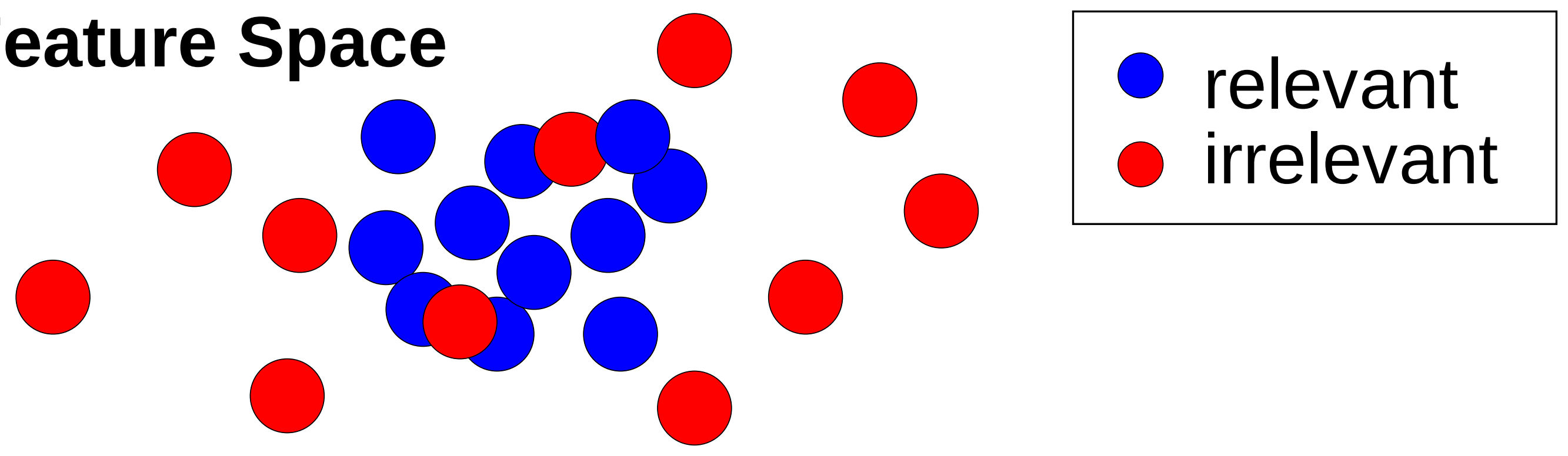
➢ Re-Ranking

$$\hat{S_Q}(x_i) = S_Q(x_i)(v(i))^{\delta}$$

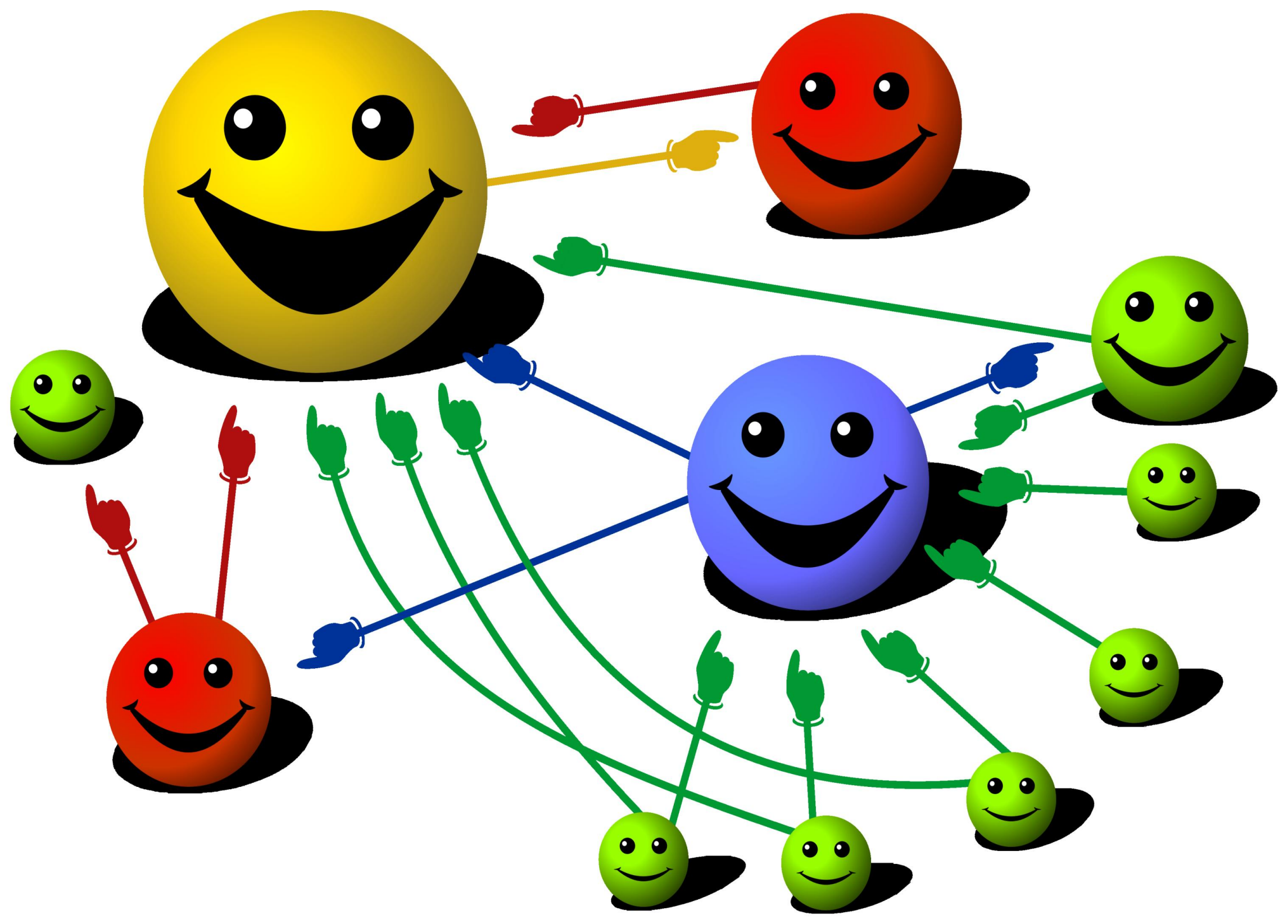
- Integrated with original relevance score

Main Idea

Acoustic Feature Space



- Relevant utterances are similar to each other in feature space.
- Utterances similar to more utterances with higher scores should be given higher relevance scores.



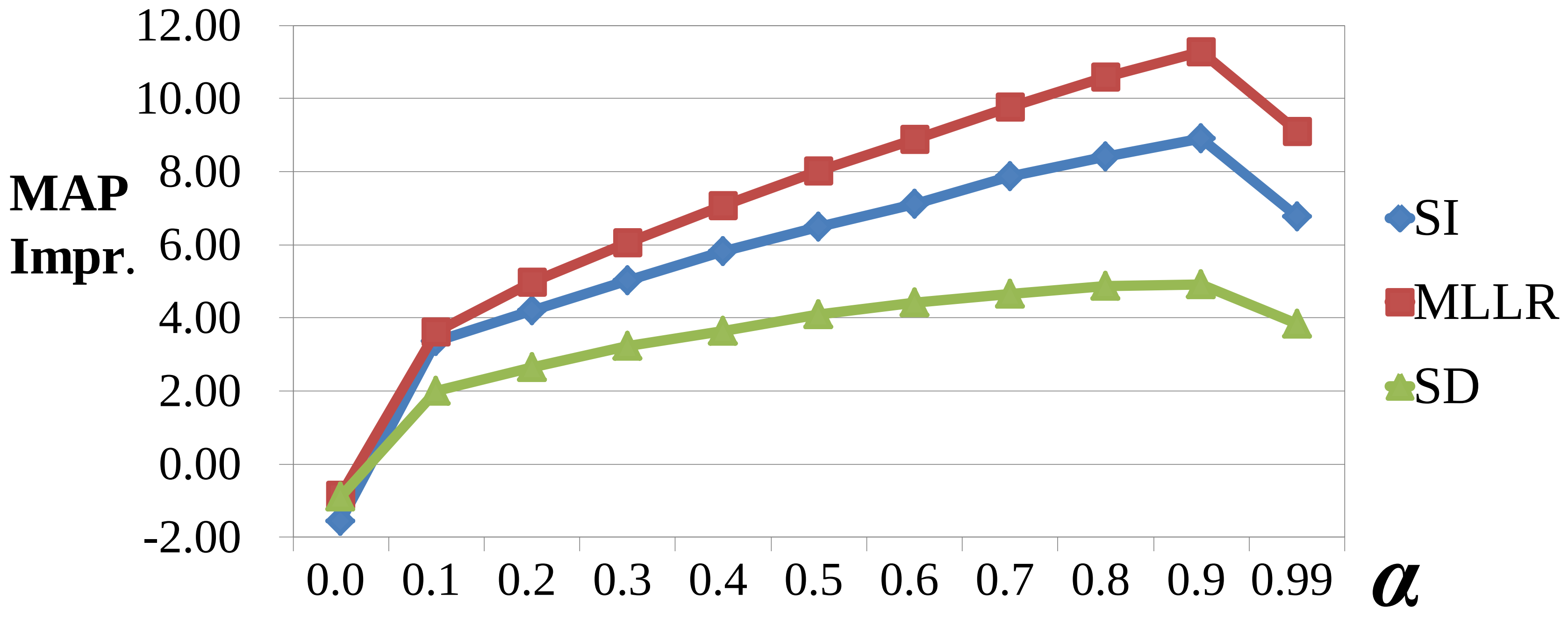
- Node: first-pass retrieved utterance
- Edge: weighted by the similarity between two utterances
- Nodes connected to more nodes with higher scores are given higher scores.
- Considering global similarity among first-pass retrieved utterances.

Experiments

- Corpus: 33 hours of course lectures (single instructor)
- Language: primarily in Mandarin Chinese
- Acoustic Model: SI, MLLR, SD

Methods	SI		MLLR		SD	
	MAP	Impr.	MAP	Impr.	MAP	Impr.
First-Pass	45.47	-	55.54	-	73.52	-
PRF	52.63	7.16	64.07	8.53	76.30	2.78
Graph	54.37	8.90	66.82	11.28	**78.44**	**4.92**
PRF + Graph	**57.75**	**12.28**	**67.38**	**11.84**	77.47	3.95

- PRF: pseudo-relevance feedback in feature space (InterSpeech 2010)
- Our approach performs better, specially for the relatively poorer acoustic models (SI and MLLR).



- The performance is optimized at ***α*** = 0.9
- Better retrieval relies primarily on global similarity.
- The graphic structure provides significant information in ranking.