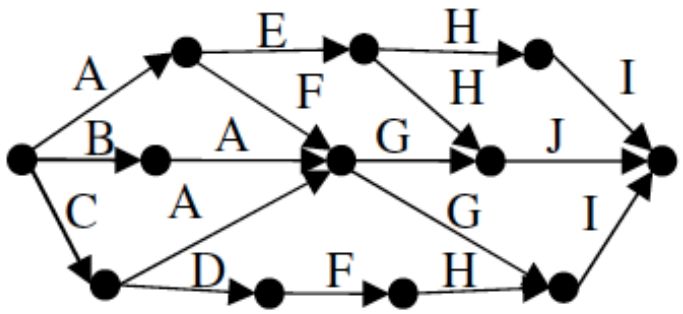
1st paper:
OPEN-VOCABULARY SPOKEN UTTERANCE RETRIEVAL USING CONFUSION NETWORKS

- Spoken utterance retrieval (SUR): retrieving short segments containing specific spoken terms or phrases from audio materials.
- Query: keywords or phrases

- Naïve approach: Using ASR (1-best-path) to index the audio to the text and then retrieving the terms from the text.

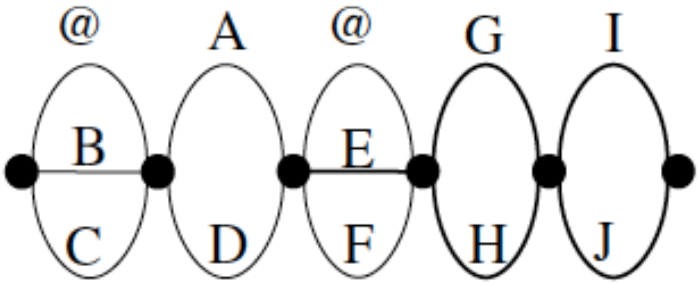
- Issue with 1-best-path ASR: recognition errors more seriously affect the retrieval performance. It can't work for OOV words.

- Alternatives:
  - retrieving from N-best path ASR.
  - using phonetic lattices specially for OOV terms. However, it dramatically increases the search space.
  - Representing lattices in most compact representation, which is called confusion network (CN).
  - Combining the word and phonetic lattices for both IV and OOV.



(a) Lattice

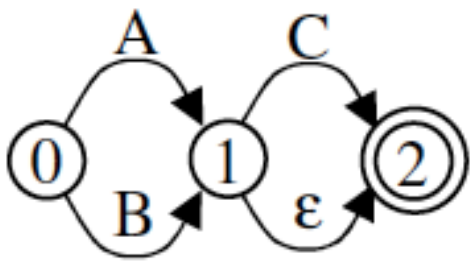
Mangu's algorithm



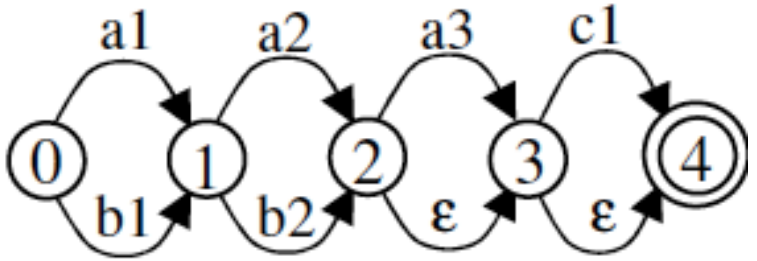
(b) Confusion network

Weights are normalized

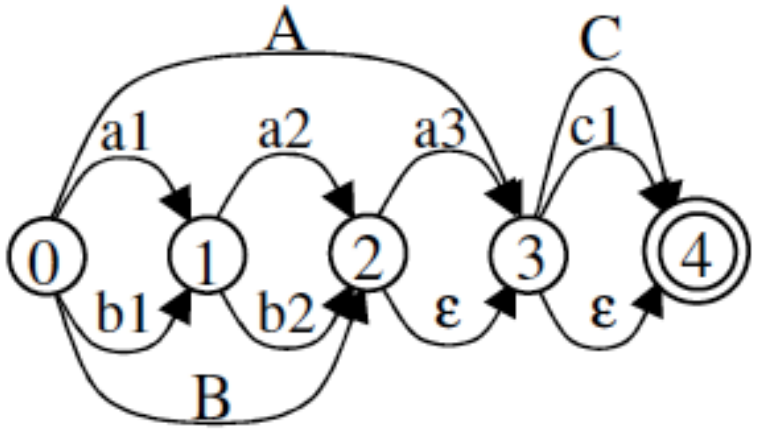
- Word-phone-combined indexing:
  - Phone-based indexing is effective for OOV, however yields lower precision for IV words. Also, the search space is an issue.
  - Combining word and sub-word indexing is effective for both IV and OOV.
  - Combining word and phone confusion networks using WFST to align two networks.
  - Pruning phone arcs that are overlapping high confidence words to reduce the table size.



(a) Word confusion network



(b) Phone confusion network



(c) Word-phone combined network

- Utterance search:
  - Representing the query in a automaton sequence.
  - Given a query, finding set of utts which probably contain the query in two steps:
    1. Finding the utts and arcs which have the label in query ignoring the proximity.
    2. Checking the proximity of arcs for each utt found in the first step. It is done by doing automaton intersection between query and utts.

- Experiments
  - MIT lecture corpus (computer science), 300 hours , language models are constructed from textbooks.
  - SUMMIT speech recognizer, vocabulary size: 16k words

  - 1st experiment: 6 hours from 4 lectures, WER = 37.2 %
  - 115 test queries (2.3 words per query in average), 13% OOV.

Table 1. Index table size and retrieval performance for the small data collection.

	1-best	WLAT	WCN	PLAT	PCN
Table size [MB]	0.8	6.7	2.4	14.0	5.0
F-score [%]	70.3	77.1	77.6	78.5	**80.2**
IV queries	73.1	80.0	80.4	79.8	**81.0**
OOV queries	0	0	0	54.1	**66.7**

- Experiments
  - MIT lecture corpus, 300 hours, language models are constructed from textbooks.
  - SUMMIT speech recognizer, vocabulary size: 16k words

  - 2st experiment: 15 hours from 4 lectures, WER = 43.9 %
  - 185 test queries (1.8 words per query in average), 13% OOV.

Table 2. Index table size and retrieval performance for the middle-size data collection.

	1-best	WLAT	WCN	PLAT	PCN
Table size [MB]	3.0	26.7	9.5	59.4	20.8
Max F-score [%]	82.5	83.8	83.8	74.9	**74.5**
IV queries	85.0	86.3	86.2	76.4	**75.9**
OOV queries	0	0	0	27.9	**38.4**

- Combined word-phone indexing

Table 3. Index table size and retrieval performance with word-phone combined indexing for the middle-size data collection.

Threshold	-	0.95	0.8
Table size [MB]	51.6	36.1	29.2
Max F-score [%]	85.1	85.0	84.8
IV queries	86.2	86.2	86.2
OOV queries	52.0	51.7	48.4

2nd paper:

BALANCING FALSE ALARMS AND HITS IN SPOKEN TERM DETECTION

- Finding a good operational point for spoken term detection (SPD) is always challenging, specially when the queries are OOV terms.

query	GAVLAK					
reference pron	G	AE	V	L	AA	K
L2S 6 best prons	G Y	AE AA AX EY	V	L	AA AY	K
index	**Candidate Hits**					
word decode	GET			LIKE		
hybrid decode	G_AE_V			L_EH_K		
phonetic lattice from hyb	G	AE	V	L	EH AE	K

- WFST based indexing is used to model phonetic confusability.

- The paper addresses reducing the false alarm rate (FAR) while increasing true hits in a STD system.

- 2 approaches for reducing the FAR by panelizing the confidence score.

- 2 approaches for increasing the hits by incorporating cache features and using phonetic confusion transducer in query representation.

- Reducing the FAR

  - OOV-detection:
    - It's used to panelize the score of an occurrence of a query-term at a time int.
    - Combination of the entropy and posterior probabilities of the sub-word unit in CN are used to provide the confidence score:

$$OOV_{scr}(\{t_j\}) = \sum_{f \in \{t_j\}} p(f|t_j)$$

    - Updating the score:

$$score_Q(\Delta_t, \gamma_o) = \begin{cases} score_Q(\Delta_t) & OOV_{scr}(\Delta_t) > 0 \\ score_Q(\Delta_t) \times \gamma_o & o/w \end{cases}$$

- Reducing the FAR
  - Query Length Normalization:
    - Incorporating a penalty term based on the length of the query term.
    - Hits with a longer duration are less likely to be false alarm.

$$score_q(\Delta_t, \gamma_L) = score_q(\Delta_t)^{\frac{\gamma_L}{\Delta_{avg}(q)}}$$

$$\gamma_L \in [0, 1]$$

$$score_q(\Delta_t, \gamma_L) = score_q(\Delta_t)^{\frac{\gamma_L}{\Delta_{avg}(q)}}$$

- Increasing hits
  - Cache features
    - Assumption: rare words tend to appear in bursts.

$$score_Q(\Delta_t, \delta) = score_Q(\Delta_t)^{1/\#hits \in \Delta_{t \pm \delta}}$$

$$\delta \in [0, 1000]$$

- Increasing hits
  - Incorporating phonetic confusions
    - Query is represented using composition of following WFST:

$$qfst = \text{bestpathN}(I(q) \circ L2S \circ P2P)$$

    - I(q): character transducer
    - L2S: letter to sound transducer
    - P2P: phonetic confusion transducer

$$score_q(\Delta_t, \gamma_L) = score_q(\Delta_t)^{\frac{\gamma_L}{\Delta_{avg}(q)}}$$

- 1st Experiment

  - 100 hours, 1290 OOVs, 5 hours DEV set
  - IBM Hybrid LVCSR, 300hours of HUB4, voc size: 83k
  - ATWV: average term weighted value

P2P-Nbest	none	10best	20best	100best
ATWV	0.342	0.368	0.384	**0.398**
%rel improv	-	7.6%	12.3%	**16.4%**

oovdet	length-norm	cache	Hits	FAs	ATWV
			9027	28472	0.398
x			8611	24378	0.399
	x		10053	25630	0.412
		x	9027	28472	0.398
	x	x	10053	25630	0.412
x	x	x	10320	35811	**0.415**

Table 4. OOVCORP Results using Automatic OOV-detector, Length-normalization, and Cache Features.

- 2nd Experiment
  - NIST 2006 STD Dev06, 3 hours , 16 OOV
  - 1107 query

oovdet	length-norm	cache	Hits	FAs	ATWV
			4752	388	0.849
x			4752	383	0.8497
	x		4845	427	0.8520
		x	4907	400	0.8551
x	x	x	5011	452	**0.8597**

Table 2. DEV06 Results using Automatic OOV-detector

Question

Vocabulary Independent Spoken Term Detection

Maider Lehr

November 8, 2012

Spoken Document Retrieval vs. Spoken Term Detection

Spoken Document Retrieval (SDR):

- Find spoken documents that are *relevant* to a given query
- Retrieval performance is quite flat with ASR WER variations in the range of 10-35%
- SDR of BN speech has been thought of as a *solved* problem

Spoken Term Detection (STD):

- Queries are usually short (1-3 words)
- Find the occurrence positions of a queried term in the spoken document

(*A survey on spoken document indexing and retrieval*. Berlin Chen, 2008)

Spoken Term Detection (Keyword spotting)

Usually 2 phases:

- Indexing:
  The speech data is automatically transcribed with an ASR and the index is created
- Searching

Automatic Speech Recognition system

ASR output in the form of:

- 1-best transcript
- lattice: acyclic directed graph
- WCN; compact representation of a lattice

The last 2 options improve the recall of the term detection

ASR transcription granularity

Word-level transcripts

- More accurate audio indexing
- Issues with the OOV terms
- Lower recall

Phone-level transcripts

- No issues with the OOV terms
- Higher recall but lower precision
- not appropriate for IV terms

Combination of word and phone level transcripts

- Word-level transcripts for IV terms in the form of WCN
- Phone-level transcripts for OOV terms in the form of 1-best transcripts

Proposed approach:

- Phrase queries with only IV terms use the word index created from the word-level transcripts
- Phrase queries with only OOV terms phone index created from the phone-level transcripts
- Phrase queries with IV and OOV terms:
  - Posting lists of the IV terms retrieved from the word index are merged with the posting lists of the OOV terms retrieved from the phonetic index.
  - The merging is done based on the timestamps stored in the posting lists.

Indexing

The index contains the following information for each unit u in a transcript D:

- begin time t of the occurrence u
- duration d of the occurrence u
  For the WCN-based indexing add:
- confidence level of occurrence u given by the posterior probability
- rank of the occurrence u w.r.t. the other candidates beginning at the same time

Search

- If the query term in ASR vocabulary:
  Use the word index to extract the posting list of the term
- If the query term not in the ASR vocabulary:
  - map the query term into phone sequence
  - extract the posting lists for each phone
  - merge the results for each phone based on the timestamps
- merge results for the IV and OOV terms based on timestamps

Ranking

- Score of the IV terms:

$$score(k, t, D) = B_{rank(k|t,D)} x Pr(k|t, D)$$

- Score of the OOV terms:

$$score(k, t_0^k, D) = 1 - \frac{\sum_{i=1}^{l} 5x(t_i^k - (t_{i-1}^k + d_{i-1}^k))}{l}$$

- Combination of scores:

$$score(Q, t_0, D) = \prod_{i=0}^{n} score(k_i, t_i, D)^{\gamma_n}$$

Experimental results: Data

Evaluation set: NIST set for STD 2006 evaluation (3hrs each set)

Corpus	BNEWS	CTS	CONFMTG
WER 1-best (WCN)	12.7	19.6	47.4

- word WCN + phone 1-best combination is only used for the BNEWS and CTS system
- phonetic transcripts too high error rates on CONFMTG
  Does it mean that precision/recall will be =0 for queries with OOV?

Experimental results

WCN vs 1-best

- Using WCN instead of 1-best improves the recall without significantly degrading the precision.
- WCN output does not bring any benefit for the CONFMTG corpus

Length of the query:

- Better performance for longer queries
- ASR is more accurate on long words

OOV vs. IV query processing

- Including the phonetic-level indexing helps with the detection of queries with OOV terms
- For queries with only OOV terms high false alarm rate

Discussion

- The timestamp information is included in the indexing
- Using the phonetic indexing only helpful for domains where the performance of the ASR is acceptable.
- Phonetic deletions should be taken into account for conversational speech
- They do not propose any solution for the OOV terms in domain with high WERs
- Their approach requires tuning several parameters