國語語音屬性偵測器之初步經驗

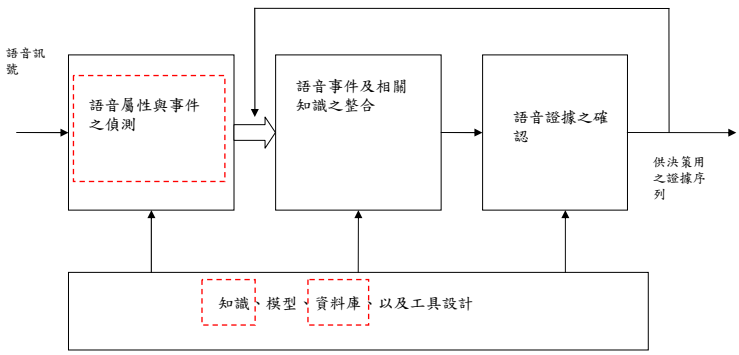
交通大學電信系 王逸如

Outline

- 前言
- 用 TIMIT 製作之英語語音屬性偵測器
- 使用國語語音屬性偵測器來偵測國語/英語語音屬性
- 國語數字串適合做國語語音屬性偵測器效能評估語料嗎?

前言

New generation ASR



Detectors in New-generation ASR

- Issues of detectors in new-generation ASR
  - What kinds of attributes, events can/need to detect?
  - What kinds of acoustic features can be used in the detectors?
  - The architectures of detectors.
- Detectors using Statistical methods
  - Labeled training data were needed.

Labeled speech data in Mandarin?

- Auto-labeling Mandarin speech data using HMM in order to get training data for detectors
  - The labeling accuracy of phones with short duration such as stops, are poor.
- Are detectors cross-language?
  - The attributes and events in speech are language independent?

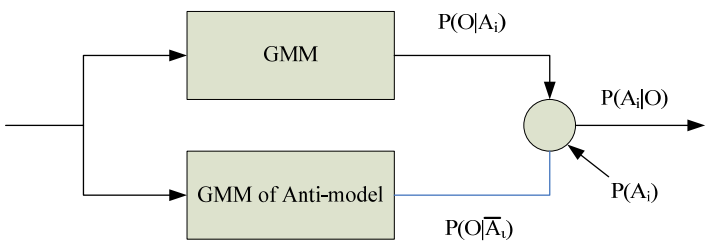
用 TIMIT 製作之英語語音屬性偵測器

- TIMIT database
  - Train : 3.8 hrs, 140,000 phones
    Test : 1.4 hrs, 50,000 phones
  - Manner: Vowel, Fricative, Stop, Nasal, Glide, Affricate
  - Position: Bilabial, Lab-dent, Dental, Alveolar, Velar, Glottal, Rhotic, Front, Central, Back

- Some statistics of TIMIT

	TIMIT Training Data				TIMIT Testing Data			
	total frames : 1,416,713				total frames : 513,526			
Manner	count	Frame number	Min Frame (10ms/ frame)	Average frame	count	Frame number	Min Frame	Average frame
Vowel	57463	549896	<1	9.57	20911	202289	1	9.67
Fricative	21424	195416	<1	9.12	7724	71036	<1	9.20
Stop	25871	106575	<1	4.12	9176	37755	<1	4.11
Nasal	14157	80454	<1	5.68	5104	29043	<1	5.69
Glide	20257	129666	<1	6.40	7822	51199	1	6.55
Silence	35877	340525	<1	9.48	12777	117734	<1	9.20
Affricate	2031	14181	2	6.98	631	4470	2	7.08

- Architectures of base detector
  - GMM based Bayesian detector



A_i and $\overline{A}_i$ is the model, anti-model of attribution *i*

- Performance of pronunciation manner detections

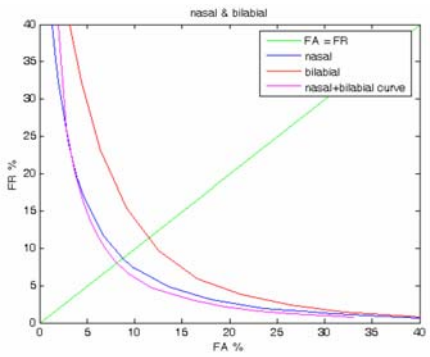
	Frame-based detector		Segment-based detector	
EER(%)	Bayesian	ANN	HMM	SEG_MCE
Vowel	12.3	9.0	1.7	1.8
Fricative	10.0	11.3	6.4	3.6
Stop	16.7	14.5	9.9	5.4
Nasal	8.7	12.2	11.2	5.4
Glide	16.3	15.9	8.0	6.1
Silence	9.7	3.7	2.1	0.8
Affricate	7.2			

- Performance of pronunciation position detectors

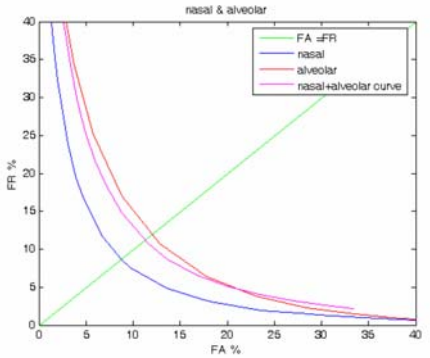
EER(%)	GMM-based Bayesian detector
Bilabial	12.2
Lab-dent	11.0
Dental	12.7
Alveolar	12.0
Velar	12.4
Glottal	18.3
Rhotic	9.4
Front	13.5
Central	17.7
Back	17.8

- Do we need Manner-position joint detectors?
- Combine the results of manner and position detectors

/m/, /em/



/n/, /en/, /nx/



使用國語語音屬性偵測器來偵測國語/英語語音屬性

- Without labeled Mandarin speech database
  - Use phone-level auto-alignment result to train the Mandarin manner detectors
  - The performance of Mandarin manner detectors for English speech data
  - The performance of Mandarin manner detectors for Mandarin speech data

- Mandarin training set
  - TCC-300 Mandarin speech database
  - Train : 23.9 hrs, 300,000 syllables
    Test : 2.4 hrs, 34,000 syllables
- Force aligned the training data using 3-state CI phone-level HMMs
- Train the GMM-based Bayesian Mandarin manner detectors

- Performance of pronunciation manner detections of Mandarin speech

	Frame-based Bayesian detector	
EER(%)	**English**	**Mandarin**
Vowel	**12.3**	**10.70**
Fricative	**10.0**	**15.7**
Stop	**16.7**	**11.5**
Nasal	**8.7**	**11.5**
Glide/Liquid	**16.3**	**9.2**
Silence	**9.7**	**8.0**
Affricate	**7.2**	**11.5**

- Compare the detecting results of TIMIT speech data using detectors trained from English/Mandarin
  - Labeling errors in Mandarin training data
  - environment miss-match

Test data : TIMIT	Frame-based detector	
EER(%)	detector trained from English	detector trained from Mandarin
Vowel	12.3	21.3
Fricative	10.0	26.1
Stop	16.7	31.0
Nasal	8.7	15.6
Glide (Liquid)	16.3	44.5 /l/
Silence	9.7	24.0
Affricate	7.2	18.5

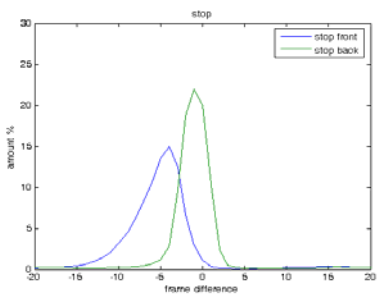
- Examples of the detection results of TIMIT-trained and TCC-trained manner detectors.

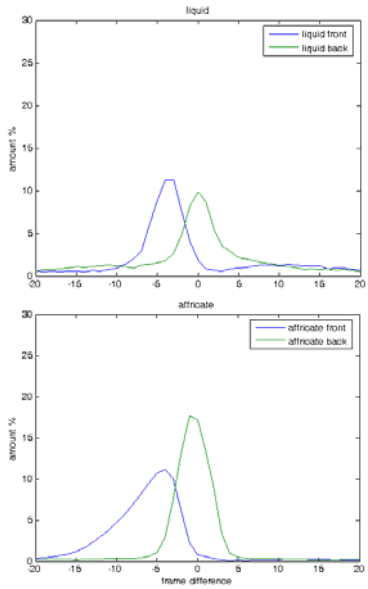
- HMM force-alignment result is poor
  - Could not find Inter-syllable silence
  - The training data of Stop, fricative, affricate, silence were poor

- Treat the GMM models in manner detectors as a 1-state HMM, they can used to force align the TCC-300 database

		Manner-based 1-state HMM		HMM	
manners	count	min frame	Average frame	min Frame	Average frame
Vowel	418337	1	8.80	3	9.77
Fricative	74276	1	8.71	3	11.17
Stop	76291	1	4.31	3	8.30
Nasal	119535	1	7.26	3	5.80
Liquid	14653	1	8.18	3	6.83
Silence	350316	0	7.53	0	4.16
Affricate	75889	1	3.88	3	10.30

- Segmentation position difference of stops, liquid, affricates





國語數字串適合做國語語音屬性偵測器效能評估語料嗎?

- Evaluation and Test set
  - To test the performance of new generation ASR?
  - Attribute-dependent test sets are needed
  - Labeled and attribute-rich database

- The manner/position attributes of Mandarin digits

	Bilabial	Lab-dent	Dental	Alveolar	Velar	Palatal	Front	Central	back
Vowel							yi, a_n	a, er, e_ng, e_n	wu, ou
Fricative			s						
Stop	b								
Nasal			n_n		ng				
Affricate						q, j			
Liquid			l						

g, k, h