MULTILINGUAL AND CROSSLINGUAL SPEECH RECOGNITION USING PHONOLOGICAL-VECTOR BASED PHONE EMBEDDINGS

Chengrui Zhu, Keyu An, Huahuan Zheng, Zhijian Ou
Speech Processing and Machine Intelligence (SPMI) Lab, Tsinghua University

IEEE ASRU 2021

Section Content

1. Motivation
2. Related work
3. Method: **JoinAP**
4. Experiments
5. Conclusion

Motivation

- There are more than 7100 languages in the world, and most of them are low-resourced languages.
- Multilingual speech recognition
  - Training data from a number of languages (seen languages) are merged to train a multilingual AM.
- Crosslingual speech recognition
  - The target language is unseen in training the multilingual AM.
  - In few-shot setting , the AM can be finetuned on limited target language data.
  - In zero-shot setting , the AM is directly used without finetuning*.

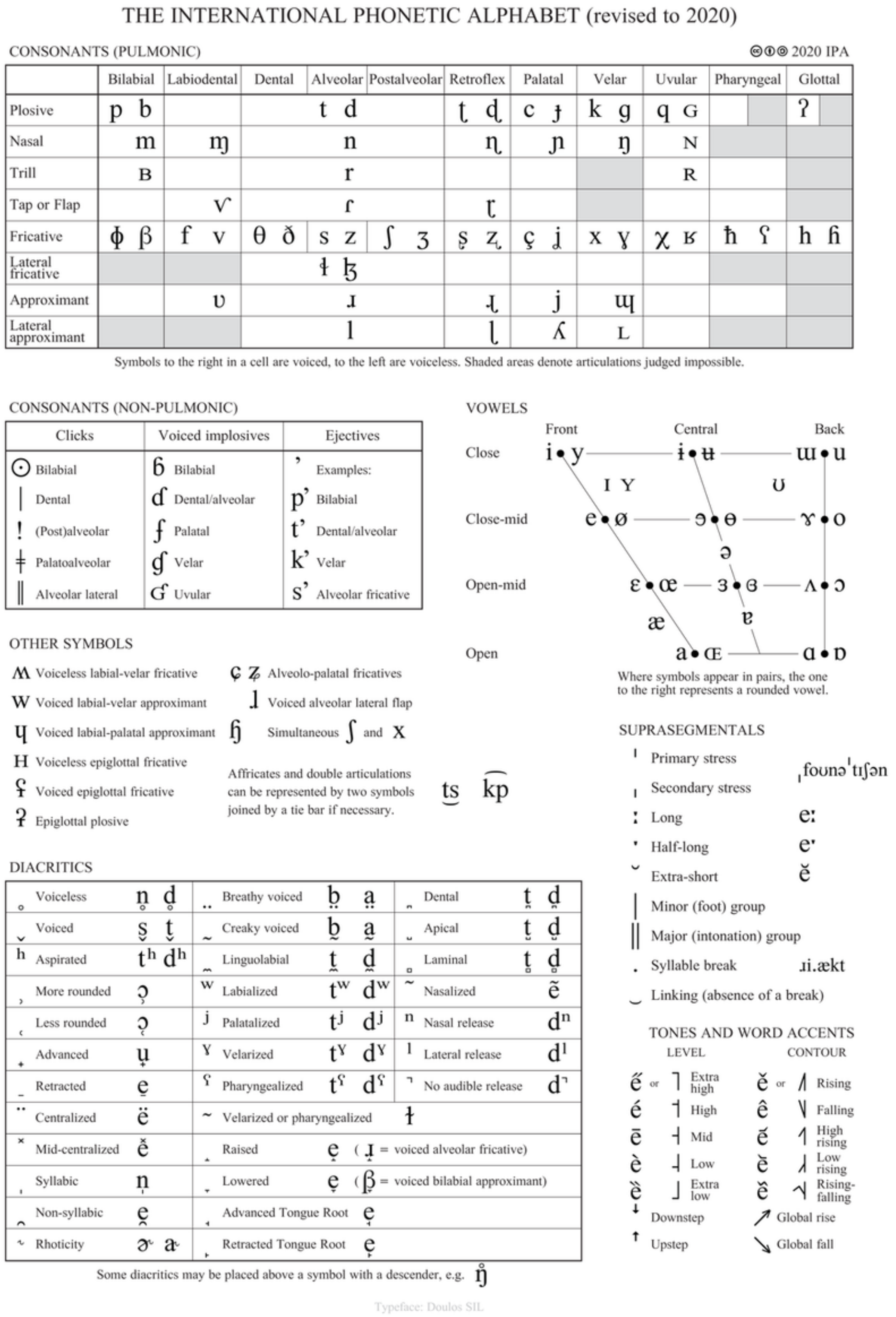
  * Suppose that text corpus from the target language are available.

Intuitively, the key to successful multilingual and crosslingual recognition is
to promote the information sharing in multilingual training
and maximize the knowledge transferring from the well trained multilingual model to the model
for recognizing the utterances in the new language.

Universal Phone Set

- International Phonetic Alphabet (IPA)

- Often phones are seen as being the “atoms” of speech. But it is now widely accepted in phonology that phones are decomposable into smaller, more fundamental units, sharable across all languages, called phonological (distinctive) features.



THE INTERNATIONAL PHONETIC ALPHABET (revised to 2020)

CONSONANTS (PULMONIC) ©️ 2020 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

Clicks		Voiced implosives		Ejectives	
ʘ	Bilabial	ɓ	Bilabial	ʼ	Examples:
ǀ	Dental	ɗ	Dental/alveolar	pʼ	Bilabial
ǃ	(Post)alveolar	ʄ	Palatal	tʼ	Dental/alveolar
ǂ	Palatoalveolar	ɠ	Velar	kʼ	Velar
ǁ	Alveolar lateral	ʛ	Uvular	sʼ	Alveolar fricative

OTHER SYMBOLS

ʍ Voiceless labial-velar fricative
w Voiced labial-velar approximant
ɥ Voiced labial-palatal approximant
ʜ Voiceless epiglottal fricative
ʢ Voiced epiglottal fricative
ʡ Epiglottal plosive
ɕ ʑ Alveolo-palatal fricatives
ɺ Voiced alveolar lateral flap
ɧ Simultaneous ʃ and x

Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary. t͡s k͡p

VOWELS

Front Central Back
Close i • y — ɨ • ʉ — ɯ • u
ɪ ʏ ʊ
Close-mid e • ø — ɘ • ɵ — ɤ • o
ə
Open-mid ɛ • œ — ɜ • ɞ — ʌ • ɔ
æ ɐ
Open a • ɶ — ɑ • ɒ

Where symbols appear in pairs, the one to the right represents a rounded vowel.

DIACRITICS

◌̥ Voiceless	n̥ d̥	◌̤ Breathy voiced	b̤ a̤	◌̪ Dental	t̪ d̪
◌̬ Voiced	s̬ t̬	◌̰ Creaky voiced	b̰ a̰	◌̺ Apical	t̺ d̺
ʰ Aspirated	tʰ dʰ	◌̼ Linguolabial	t̼ d̼	◌̻ Laminal	t̻ d̻
◌̹ More rounded	ɔ̹	ʷ Labialized	tʷ dʷ	◌̃ Nasalized	ẽ
◌̜ Less rounded	ɔ̜	ʲ Palatalized	tʲ dʲ	ⁿ Nasal release	dⁿ
◌̟ Advanced	u̟	ˠ Velarized	tˠ dˠ	ˡ Lateral release	dˡ
◌̠ Retracted	e̠	ˤ Pharyngealized	tˤ dˤ	̚ No audible release	d̚
◌̈ Centralized	ë	◌̴ Velarized or pharyngealized	ɫ		
◌̽ Mid-centralized	e̽	◌̝ Raised	e̝ (ɹ̝ = voiced alveolar fricative)		
◌̩ Syllabic	n̩	◌̞ Lowered	e̞ (β̞ = voiced bilabial approximant)		
◌̯ Non-syllabic	e̯	◌̘ Advanced Tongue Root	e̘		
˞ Rhoticity	ɚ a˞	◌̙ Retracted Tongue Root	e̙		

Some diacritics may be placed above a symbol with a descender, e.g. ŋ̊

SUPRASEGMENTALS

ˈ Primary stress ˌfoʊnəˈtɪʃən
ˌ Secondary stress
ː Long eː
ˑ Half-long eˑ
◌̆ Extra-short ĕ
| Minor (foot) group
‖ Major (intonation) group
. Syllable break ɹi.ækt
‿ Linking (absence of a break)

TONES AND WORD ACCENTS

LEVEL
e̋ or ˥ Extra high
é ˦ High
ē ˧ Mid
è ˨ Low
ȅ ˩ Extra low
ꜜ Downstep
ꜛ Upstep

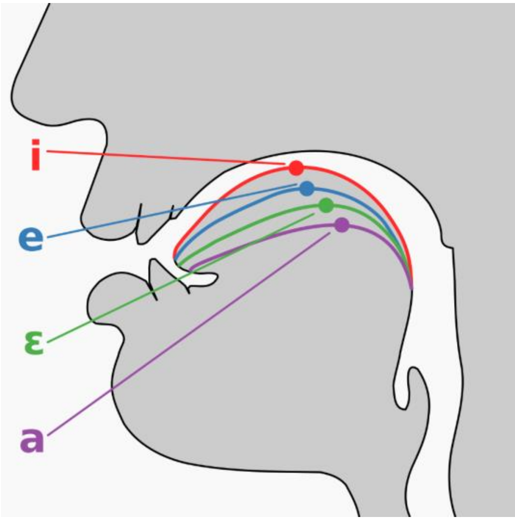
CONTOUR
ě or ˩˥ Rising
ê ˥˩ Falling
e᷄ ˦˥ High rising
e᷅ ˩˨ Low rising
e᷈ ˧˦˧ Rising-falling
↗ Global rise
↘ Global fall

Typeface: Doulos SIL

Phonological features

Describe phones by phonological features

- Vowels
  - vowel height
  - vowel backness
- Consonants
  - Place of articulation
  - Manner of articulation



Phonological feature	d	ɛ	ð	ə	i	ʥ	k^j
syllabic	-	+	-	+	+	-	-
sonorant	-	+	-	+	+	-	-
consonantal	+	-	+	-	-	+	+
continuant	-	+	+	+	+	-	-
delayed release	-	-	-	-	-	+	-
lateral	-	-	-	-	-	-	-
nasal	-	-	-	-	-	-	-
strident	0	0	0	0	0	0	0
voice	+	+	+	+	+	+	-
spread glottis	-	-	-	-	-	-	-
constricted glottis	-	-	-	-	-	-	-
anterior	+	0	+	0	0	-	-
coronal	+	-	+	-	-	+	-
distributed labial	-	0	+	0	0	+	0
labial	-	-	-	-	-	-	-
high	-	-	-	-	+	+	+
low	-	-	-	-	-	-	-
back	-	-	-	+	-	-	-
round	-	-	-	-	-	-	-
velaric	-	-	-	-	-	-	-
tense	0	-	0	-	+	0	0
long	-	-	-	-	-	-	-
hitone	0	0	0	0	0	0	0
hireg	0	0	0	0	0	0	0

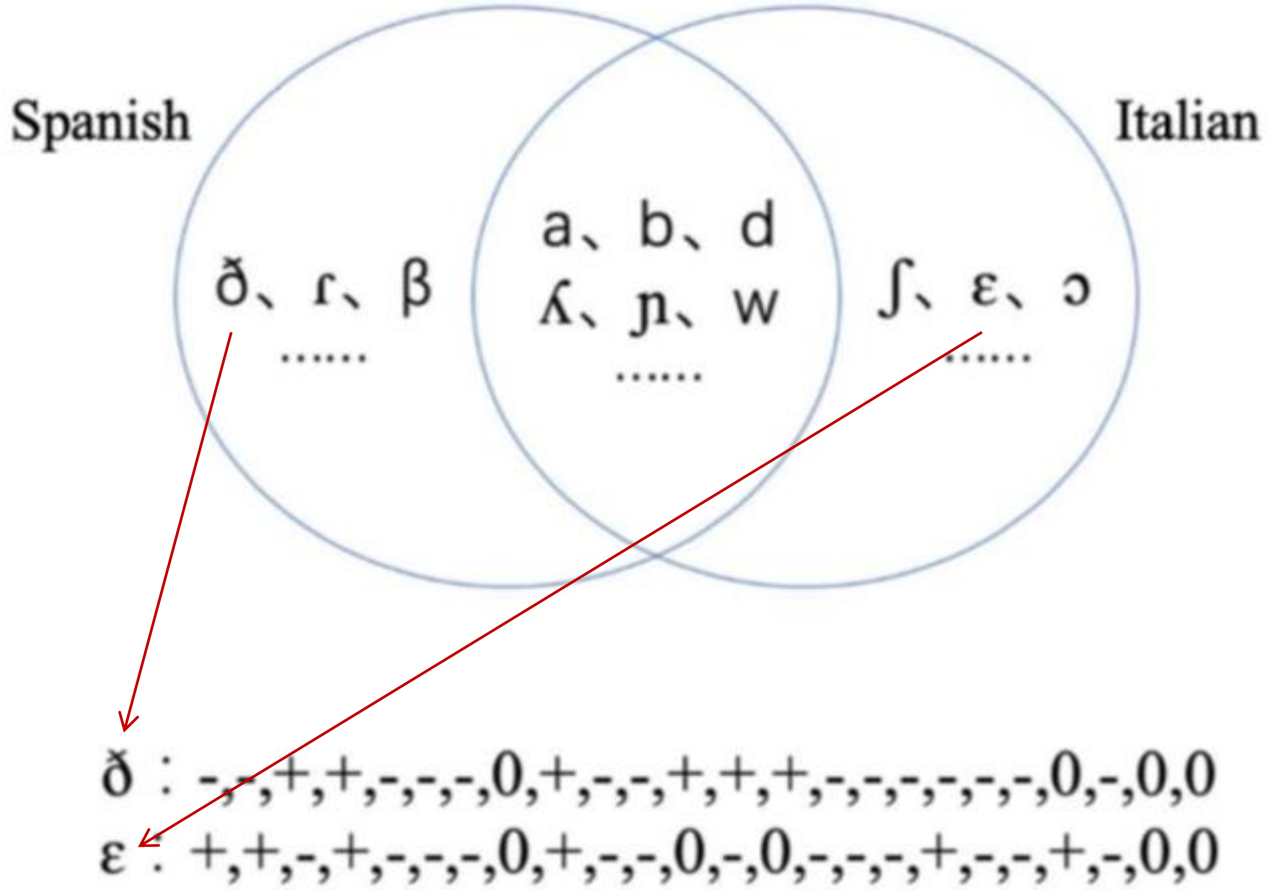
Phonological features: micro-decomposition of phones

- Like atoms could be split into nucleus and electrons, phones can be expressed by phonological features.

Matter	Speech
Atoms	Phones
Periodic table of elements	IPA table
Nucleus, electrons	Phonological features

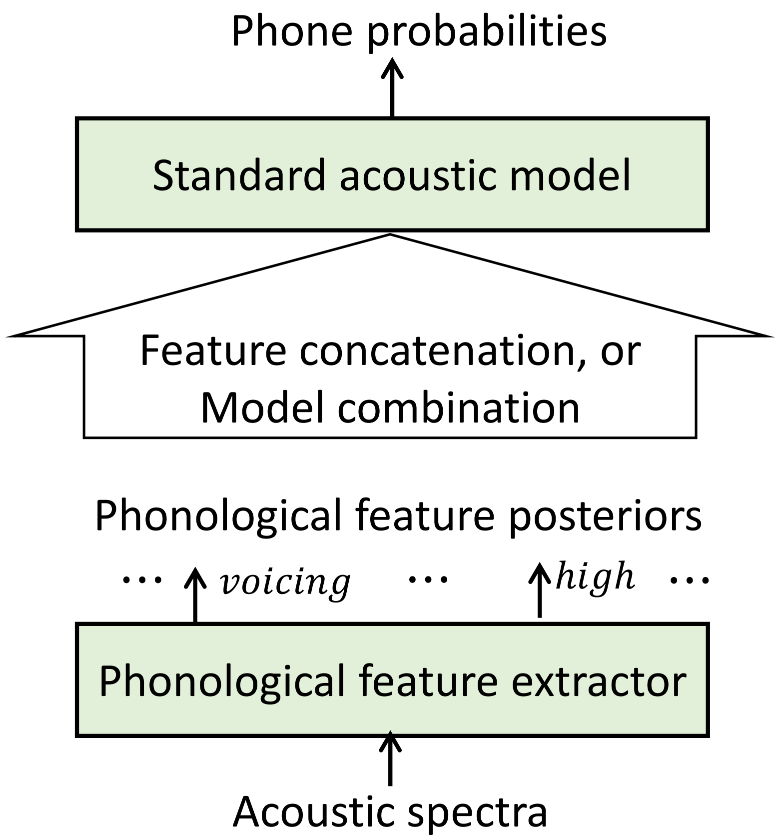
Phonological features: promote information sharing

- Even language-specific phones are connected by using phonological features.



Related work

- Phonological features(PFs) have been applied in multilingual and crosslingual ASR
- Previous studies generally take a bottom-up approach, and suffer from:
  - The acoustic-to-PF extraction in a bottom-up way is itself difficult.
  - Do not provide a principled model to calculate the phone probabilities for unseen phones from the new language towards zero-shot crosslingual recognition.



From phonological features to phonological-vector

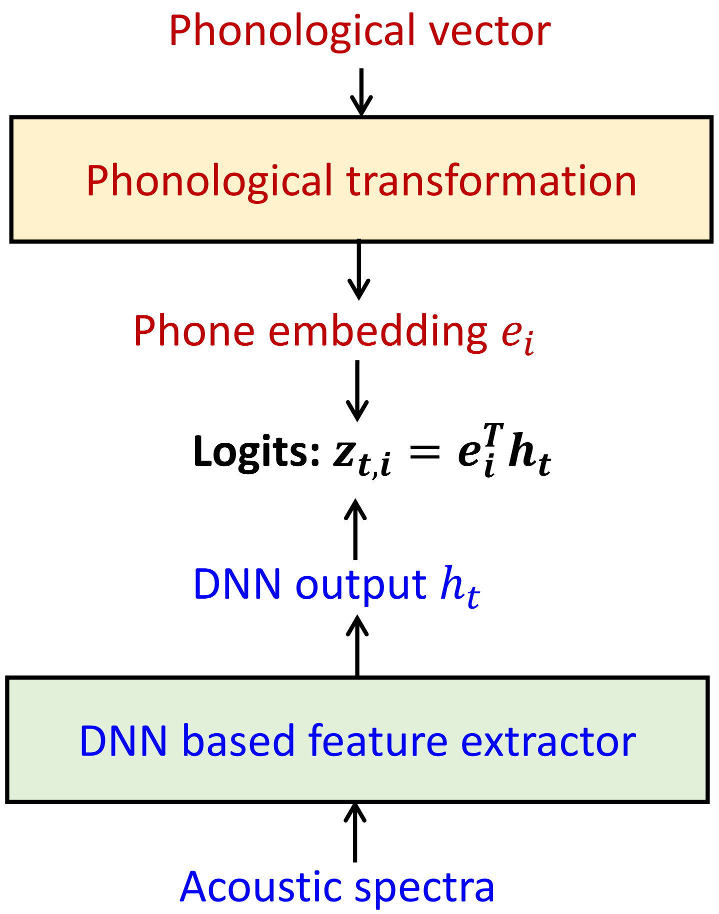
- Phonological-vector
  - Encode each phonological feature by a 2-bit binary vector. (24PFs -> 48bits)

+	-	0
10	01	00

  - Plus 3 bits to indicate <blk>, <spn>, <nsn>
  - Phonological-vector: Total 51 bits

Joining of Acoustics and Phonology (JoinAP)

- The JoinAP method
  - DNN based acoustic feature extraction (bottom-up) and phonology driven phone embedding (top-down) are joined to calculate the **logits**.



- JoinAP-Linear
  - Linear transformation of phonological-vector p_i to define the embedding vector for phone i:

$$e_i = Ap_i \in \mathbb{R}^H$$

- JoinAP-Nonlinear
  - Apply nonlinear transformation, multilayered neural networks:

$$e_i = A_2\sigma(A_1p_i) \in \mathbb{R}^H$$

Experiments

- Train multilingual AM on German, French, Spanish and Italian.
- Zero-shot and few-shot crosslingual ASR on Polish and Mandarin.
- Employ Phonetisaurus G2P to generate IPA lexicons
- Use CTC-CRF based ASR toolkit, CAT
  - Acoustic model: 3 layer VGGBLSTM with 1024 hidden dim
  - Adam optimizer: with an initial learning rate of 0.001, decreased to 1/10 until less than 0.00001
  - Dropout 0.5

Language	Corpora	#Phones	Train	Dev	Test
German	CommonVoice	40	639.4	24.7	25.1
French	CommonVoice	57	465.2	21.9	23.0
Spanish	CommonVoice	30	246.4	24.9	25.6
Italian	CommonVoice	33	89.3	19.7	20.8
Polish	CommonVoice	46	93.2	5.2	6.1
Mandarin	AISHELL-1	96	150.9	18.1	10.0

Experiments

- Multilingual experiments

Language	Flat-Phone monolingual	Flat-Phone w/o finetuning	Flat-Phone finetuning	JoinAP-Linear w/o finetuning	JoinAP-Linear finetuning	JoinAP-Nonlinear w/o finetuning	JoinAP-Nonlinear finetuning
German	13.09	14.36	12.42	13.72	12.45	13.97	12.64
French	18.96	22.73	18.91	22.73	19.54	22.88	19.62
Spanish	15.11	13.93	13.06	13.93	13.19	14.10	13.26
Italian	24.57	25.97	21.77	25.85	21.70	24.06	20.29
Average	17.93	19.25	16.54	19.06	16.72	18.75	16.45

- Language-degree of a phone: how many languages a phone appears

Language \ Language-degree	4	3	2	1
German	18	6	8	8
French	18	6	7	26
Spanish	18	4	1	7
Italian	18	5	4	6

On average, both JoinAP-Nonlinear and JoinAP-Linear perform better than Flat-Phone, and JoinAP-Nonlinear is the strongest.

Experiments

- Crosslingual experiments
  - Polish:

#Finetune	Flat-Phone	JoinAP-Linear	JoinAP-Nonlinear
0	33.15	35.73	31.80
10 minutes	8.70	7.50	8.10

  - Mandarin:

#Finetune	Flat-Phone	JoinAP-Linear	JoinAP-Nonlinear
0	97.10	89.51	88.41
1 hour	25.39	25.21	24.86

  - Statistics about Polish and Mandarin:

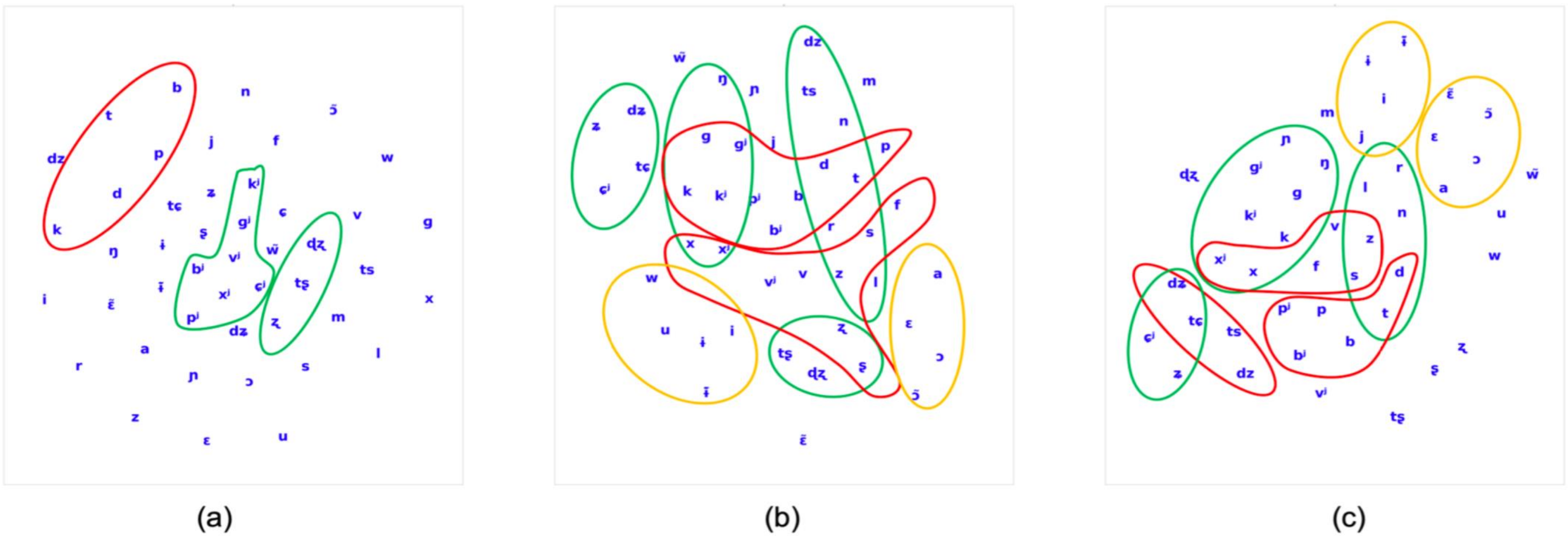
Language	#Phones	#Unseen phones
Polish	46	18
Mandarin	96	79

On average, both JoinAP-Nonlinear and JoinAP-Linear perform better than Flat-Phone, and JoinAP-Nonlinear is the strongest.

Experiments

- t-SNE map of Polish phone embeddings

  (obtained from un-finetuned multilingual models)



(a) Flat phone embeddings, (b) JoinAP-Linear phone embeddings, (c) JoinAP- Nonlinear phone embeddings.

Consonants with the same manner of articulation

Consonants with the same place of articulation

Vowel with similar height

Conclusion

- In the multilingual and crosslingual experiments, JoinAP-Nonlinear generally performs better than JoinAP-Linear and the traditional flat-phone method on average. The improvements for target language depend on its data amount and language-degree.

- Our JoinAP method provides a principled, data-efficient approach to multilingual and crosslingual speech recognition.

- Promising directions: exploring DNN based phonological transformation, and pretraining over increasing number of languages.