An Investigation On Training Deep Neural Networks Using Probabilistic Transcriptions

Amit Das, Mark Hasegawa-Johnson
Dept. of ECE & Beckman Institute, University of Illinois, USA

Beckman Institute for Advanced Science and Technology
University of Illinois at Urbana-Champaign

1. Objectives

How do we train DNNs (Deep Neural Networks) using **noisy transcriptions** collected from **non-native crowd workers** on the web? These workers do not speak the native/target language. Hence, the transcriptions they provide are prone to labeling errors. Standard DNN training using such transcriptions do not yield significant improvements over GMM-HMM systems.

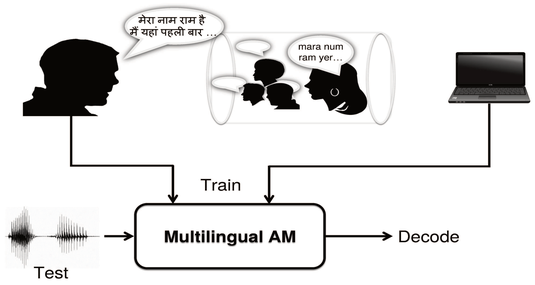
Main Contributions:

- ✓ Investigate DNN training methods effective for training using non-native crowd transcriptions.
- ✓ Achieved consistent and absolute improvement in PERs (phone error rates) in the range 1.3-6.2% over GMM-HMM.

2. Motivation

- Rich resourced languages (e.g. English) have >100 hours of labeled data and >100k entries in their lexicon. This is the typical requirement to build good ASR systems.
- For under-resourced languages, it is possible to find speech waveforms and text data but hard to find corresponding labels for those waveforms.
- In the absence of native transcripts, we resort to collecting *approximate* transcripts collected from non-native crowd workers. And, train ASR systems using these transcripts.

3. Scenario



- We do not have natively transcribed labeled data in the target language **L**. Transcripts from native transcribers are called **DTs** (**deterministic transcripts**).
- We have non-natively transcribed labeled data from crowd workers who neither speak nor have any familiarity with **L**. Hence, these transcripts are prone to labeling errors. Transcripts from crowd workers are called **PTs** (**probabilistic transcripts**).

4. DT vs PT

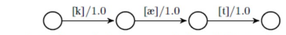


Figure 1: A deterministic transcription (DT) for the word *cat*.

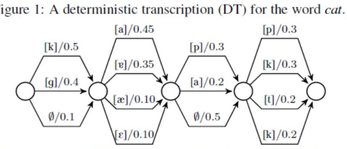


Figure 2: A probabilistic transcription (PT) for the word *cat*.

	DT	PT
Transcribers	Native	Crowdworkers
Transcription Structure	Single stream	Lattice
Probability	1.0	$(0, 1]$
Label Noise	Low	High
Availability	Difficult	Easy

5. Corpus

Table 1: SBS Multilingual Corpus ($\approx$ 40 min per language)

Language	Utterances		Phones
	Train	Test	
Swahili (SW)	463	123	53
Hungarian (HG)	459	117	70
Cantonese (CA)	544	148	37
Mandarin (MD)	467	113	57
Arabic (AR)	468	112	51
Urdu (UR)	385	94	45
All	-	-	83

- We pick a language as the test/target language, **L**, from the above table.
- Then training data for **L** are:
  - PTs in **L**. But we do not include DTs in **L**.
  - DTs in all the other languages ($\neq$ **L**).
  - Any unlabeled data that might be available in **L**.

Example: Swahili as a test langauge

Table 2: Training data for Swahili

Language	Transcript Type	Amount of Data
SW	PT	40 min
HG + CA + MD + $\cdots$	DT	200 min
AR + UR		(40 min $\times$ 5 lang)
SW	Unlabeled	5 hrs

Note: DTs in Swahili are never used for training.

6. Training DNN Using Probabilistic Transcriptions

- Standard DNN cross-entropy training not significantly better (or sometimes worse) than GMM-HMM when trained using PTs.
- **Problem:** Discriminative training is sensitive to noise.
- Example of a frame containing 'ae': (green = true label, red = incorrect label)
  - DT: Train using 1-hot labels $\Rightarrow$ [1.0 æ]
  - PT: Train using soft labels $\Rightarrow$ [0.35 a, 0.45 ɐ, 0.1 æ, 0.1 ɛ]
- **Solution: Transfer knowledge from DT to PT using Multi-task Learning (MTL).**
  - Train PT and DT using two different softmax layers. Or, train using PT, DT and ST (self-training transcripts) in three different softmax layers.
  - DT and high confidence STs help fix the errors in hidden layers. Obtain better feature separation in the hidden layers.

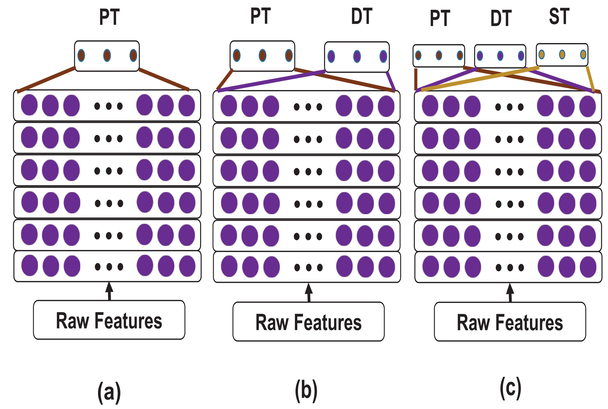


Figure 1: (a) DNN-1 (No MTL), (b) DNN-2 (MTL), (c) DNN-3 (MTL)

7. Results

Best Case PER

- ✓ DTs available in the target language (Oracle scenario); Monolingual training.

Worst Case PER

- × DTs not available in the target language.
- × PTs not available in the target language.
- ✓ DTs available in other languages; Multilingual training.

Table 3: MONO (Dev set PER inside parentheses)

Lang	PER (%)	
	HMM	DNN
SW	35.63 (47.00)	34.18 (39.49)
HG	38.72 (40.33)	35.62 (37.32)
MD	31.80 (26.14)	28.26 (25.16)

Table 4: MULTI (Dev set PER inside parentheses)

Lang	PER (%)		
	HMM	DNN	# Senones
SW	65.73 (67.58)	61.17 (63.12)	1003
HG	67.55 (68.50)	63.25 (63.65)	1012
MD	71.09 (69.10)	64.68 (63.84)	994

PER of Proposed DNNs

- × DTs not available in the target language.
- ✓ PTs available in the target language.
- ✓ DTs available in other languages.

Table 5: Comparison of PERs of baseline (MULTI, MAP HMM, DNN-1) vs proposed systems (DNN-2, DNN-3). Absolute improvement in PER over MAP HMM in parentheses.

Lang	PER (%)					
	MULTI (Worst)	MAP HMM	DNN-1	DNN-2	DNN-3	MONO (Best)
SW	65.73	44.77	45.14 (**-0.37**)	43.03 (**1.74**)	43.50 (1.27)	35.63
HG	67.55	56.85	56.13 (**0.72**)	55.53 (**1.32**)	55.69 (1.16)	38.72
MD	71.09	59.23	54.95 (4.28)	53.70 (5.53)	53.05 (**6.18**)	31.80

8. Conclusions

- Proposed DNN-2/DNN-3 systems achieved absolute improvement in PERs in the range 1.3-6.2% over GMM-HMM systems consistently.
- They are able to close between 28% and 67% (relative) of the gap between MULTI and MONO systems. Thus, PTs are between one and two thirds as useful as DTs.
- Crowdsourced PTs are not as useful as DTs. The gap between MONO and DNN-2/DNN-3 is still large.

References

[1] M. Hasegawa-Johnson *et al.*, "ASR for Under-Resourced Languages from Probabilistic Transcription," IEEE Trans. Audio Speech & Lang. Proc. (under review).