Mismatched Crowdsourcing: Mining Latent Skills to Acquire Speech Transcriptions

Mark Hasegawa-Johnson
University of Illinois
Urbana-Champaign, USA
jhasegaw@illinois.edu

Preethi Jyothi
IIT Bombay
Mumbai, India
pjyothi@cse.iitb.ac.in

Wenda Chen
I^{2}R, A*STAR
Singapore
chen-w@i2r.a-star.edu.sg

Van Hai Do
Thuyloi University
Vietnam
haidv@tlu.edu.vn

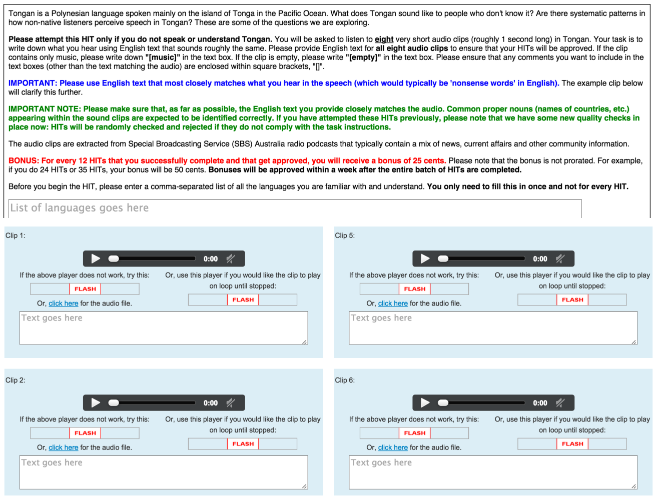
***Abstract*—Automatic speech recognition (ASR) converts audio to text. ASR is usually trained using a large quantity of labeled data, i.e., audio with text transcription. In many languages, however, text transcription is hard to find, e.g., in both Hokkien and Dinka, we found native speakers who had received all their primary education in some other language, and who therefore had difficulty writing in their own language. Fortunately, speech in every language is produced by human mouths, and designed to be interpreted by human ears. Speakers of a majority language (English, say, or Mandarin Chinese) are therefore able to make some sense of even the strangest language (Zulu, say, or Cantonese): language-unique distinctions are mostly lost, but universal distinctions such as consonant versus vowel are, for the most part, correctly transmitted. We can decode such mismatched transcripts using an information-theoretic decoder, resulting in a low-entropy probability distribution over the possible native-language transcriptions. Mismatched transcripts can be used to train ASR. Combining ten hours of mismatched transcripts with 12-48 minutes of native transcripts, if available, results in lower phone error rate. On the other hand, if we don't even know the native phoneme inventory, mismatched transcripts in two or more annotation languages can be used to infer the native phoneme inventory (with entropy depending on the distinctive feature inventory of the annotation languages).**

I. Introduction

ASR is, of course, the most difficult of all machine learning problems, because there is no such thing as a labeled training example. It is very easy for a human to listen to a sentence in her own native language, and to write down the words that were spoken, but it is very difficult for her to identify which of the speech sounds correspond to which of the short-time Fourier transform frames in the utterance. For this reason, all modern ASRs are trained using semi-supervised methods based either on the Baum-Welch algorithm [1] or based on the segmental K-means algorithm [2]. Typically we ask native speakers of the target language to transcribe about 200 hours of audio from the language of interest, and write down the words contained therein. Orthography is then converted to pronunciation using a grapheme-to-phoneme transducer (a difficult process only in English, Urdu, and Arabic), then Baum-Welch or segmental K-means are used to align the pronunciation with the audio, and the resulting alignments are used to train the classifiers (neural nets or Gaussian mixtures); the last two steps are iterated toward a local optimum.

In order to train an ASR in Swahili, we (1) downloaded a few hours of podcast audio in Swahili, (2) advertised at the University of Illinois to find native speakers of Swahili, (3) hired the person who responded to transcribe the audio for us. The problems with this methodology are: (1) it is only possible to create ASR in a language for which you have a native informant, (2) the quality of your ASR depends on the number of hours of transcribed audio you have, i.e., it depends on the number of hours your informant is willing and able to commit to the project. Transcription is tedious work, requiring an extraordinary ability to focus. In English and Mandarin, it is possible to hire transcribers who can transcribe an hour of audio in six hours of labor ($6 \times RT$). When we hired professional English-to-Vietnamese transcribers to transcribe Vietnamese audio for us, we set $20 \times RT$ as the minimum requirement, and were forced to lay off two out of seven workers for failing to meet that target. The problem is more acute in under-resourced languages. Our Dinka native audio corpus was transcribed by a native speaker of Dinka who received all of his formal education in English; he can transcribe English at roughly $10 \times RT$, but transcribing Dinka (his mother tongue) takes him about three times as long. Singapore Hokkien is a language with no native orthography; in order to acquire native-speaker transcriptions of Hokkien, our collaborators had to first invent an orthography for the language [3], [4], hire a native speaker, and then teach the native speaker an orthography with which to write her own mother tongue. In the most extreme case (but also the most common), it may be completely impossible to find a native speaker of the target language. The experiments described in [5], for example, attempted to develop ASR in seventy languages. We advertised on a large multi-national university campus, seeking native speakers of any of those seventy languages; native speakers of ten languages responded, of whom only six were able to completely transcribe one hour of audio.

When the target language has no skilled transcribers for hire, we choose, instead, to hire a transcription in whatever language the good transcribers are best able to understand. There are certain speech events that everyone can hear, regardless of whether or not they speak the language. For example, usually everyone can hear the differences among stop consonants, fricative consonants, nasal consonants, and vowels (most of the time); everyone can hear the difference between consonants made with the lips, the tongue tip, and parts of the tongue



Tongan is a Polynesian language spoken mainly on the island of Tonga in the Pacific Ocean. What does Tongan sound like to people who don't know it? Are there systematic patterns in how non-native listeners perceive speech in Tongan? These are some of the questions we are exploring.

Please attempt this HIT only if you do not speak or understand Tongan. You will be asked to listen to **eight** very short audio clips (roughly 1 second long) in Tongan. Your task is to write down what you hear using English text that sounds roughly the same. Please provide English text for **all eight audio clips** to ensure that your HITs will be approved. If the clip contains only music, please write down **"[music]"** in the text box. If the clip is empty, please write **"[empty]"** in the text box. Please ensure that any comments you want to include in the text boxes (other than the text matching the audio) are enclosed within square brackets, "[]".

IMPORTANT: Please use English text that most closely matches what you hear in the speech (which would typically be 'nonsense words' in English). The example clip below will clarify this further.

IMPORTANT NOTE: Please make sure that, as far as possible, the English text you provide closely matches the audio. Common proper nouns (names of countries, etc.) appearing within the sound clips are expected to be identified correctly. If you have attempted these HITs previously, please note that we have some new quality checks in place now: HITs will be randomly checked and rejected if they do not comply with the task instructions.

The audio clips are extracted from Special Broadcasting Service (SBS) Australia radio podcasts that typically contain a mix of news, current affairs and other community information.

BONUS: For every 12 HITs that you successfully complete and that get approved, you will receive a bonus of 25 cents. Please note that the bonus is not prorated. For example, if you do 24 HITs or 35 HITs, your bonus will be 50 cents. **Bonuses will be approved within a week after the entire batch of HITs are completed.**

Before you begin the HIT, please enter a comma-separated list of all the languages you are familiar with and understand. **You only need to fill this in once and not for every HIT.**

List of languages goes here

Clip 1:

0:00

If the above player does not work, try this: FLASH

Or, click here for the audio file.

Or, use this player if you would like the clip to play on loop until stopped: FLASH

Text goes here

Clip 5:

0:00

If the above player does not work, try this: FLASH

Or, click here for the audio file.

Or, use this player if you would like the clip to play on loop until stopped: FLASH

Text goes here

Clip 2:

0:00

If the above player does not work, try this: FLASH

Or, click here for the audio file.

Or, use this player if you would like the clip to play on loop until stopped: FLASH

Text goes here

Clip 6:

0:00

If the above player does not work, try this: FLASH

Or, click here for the audio file.

Or, use this player if you would like the clip to play on loop until stopped: FLASH

Text goes here

Fig. 1. The mismatched crowdsourcing task: Crowd workers are asked to listen to audio in a language they don't know, and write down what it sounds like, using nonsense syllables in English.

TABLE I
MISMATCHED TRANSCRIPTS OF THE SWAHILI PHRASE "-ENDA PAMOJA NA," TRANSCRIBED BY TEN DIFFERENT CROWD WORKERS. THESE ARE CALLED MISMATCHED TRANSCRIPTS BECAUSE THEY ARE PRODUCED BY PEOPLE WHO DON'T UNDERSTAND THE LANGUAGE THEY ARE TRANSCRIBING.

e n d a p o m o n y a m A	m e m b e r p a m o h y a
e n d a c u m o y a h	e n d a a m o y a a
e n d a p a m o y a n a	e n d a p a m o l y a
e n d a p u h m o h y a n A	e b a p u m u l i e u h
e n n a p a m U y a n A	d a p a m o j e r a

behind the tip (most of the time). In [6] we proposed the crowdsourcing task shown in Fig. 1. Crowd workers are asked to listen to an audio clip in a language they don't know, to interpret the audio as nonsense speech in English, and to "Please use English nonsense syllables that most closely match what you hear in the speech." Table I shows, as an example, transcriptions of the Swahili phrase "-enda pamoja na" as recorded by ten different crowd workers, none of whom know Swahili.

If you transcribe a language you don't know, then you will make mistakes. We model the transcriber as a noisy channel: Swahili phonemes enter the noisy channel, and English letters come out [7]. A noise-free channel would generate a perfect transcript of the audio; a uniformly random channel would generate letter sequences uniformly at random. A mismatched transcriber generates letter sequences at random, but not uniformly at random. Rather, if N different transcribers listen to the same sequence of Swahili phonemes X, then the letter sequence generated by the i^{th} such transcriber, Y_i, is generated according to some probability mass function $p(Y_i|X)$. It is reasonable to assume that the substitution probabilities (the "channel") depend only on the native language of the transcriber, and not on any other quirks of the individual; under that assumption, it is possible to estimate $p(Y|X)$ by measuring the substitution error frequencies of many English-speaking transcribers. Similarly, if the transcribers perform their tasks independently, it is reasonable to assume that the transcriptions Y_i are independent given knowledge of X. Under these assumptions, we can find the most probable phone sequence X^* as:

$$X^* = \operatorname{argmax} p(X) \prod_{i=1}^{N} p(Y_i|X) \tag{1}$$

This paper considers two problems with the mismatched crowdsourcing paradigm. First (Sec. II), we need to estimate the channel: we need to find some way to estimate $p(Y|X)$ for a certain pool of transcribers. Second, it is necessary to justify the cost of mismatched crowdsourcing, and for that purpose, it is useful to measure its utility (Sec. III) and to optimize the resulting cost-benefit tradeoff (Sec. IV). Sec. V concludes.

II. Modeling the Noise: Second-Language Speech Perception

The misperception of non-native phonemes is a standard topic of research in the field of Second-Language Speech Perception, and knowledge from that field can guide us in the design of a useful model (Sec. II-A). One of the benefits of mismatched crowdsourcing is that it provides a new method of acquiring data with which to train Second-Language Speech Perception models; methods for learning the parameters of such a model from experimental data are summarized in Sec. II-B.

A. The Perceptual Assimilation Model

Suppose that the spoken phoneme string $X = [x_1, \ldots, x_J]$ is transcribed using the written phoneme string $Y = [y_1, \ldots, y_K]$. This subsection will use Y to denote a sequence of English phonemes, rather than a sequence of English orthographic symbols; methods exist that can map from one to the other with a reasonably low probability of error.

The Perceptual Assimilation Model [8] proposes that a listener's ability to distinguish two foreign phones, x_1 and x_2, is determined by the native phones y_1 and y_2 to which they are perceptually most similar:

$$y_1 = \text{argmin}_{y \in \mathcal{Y}} D(y, x_1) \tag{2}$$

$$y_2 = \text{argmin}_{y \in \mathcal{Y}} D(y, x_2) \tag{3}$$

where $\mathcal{Y}$ means the listener's native phoneme set, and $D(x, y)$ means the perceptual distance between native phoneme y and foreign phoneme x.

In second-language speech perceptual papers, the most common way to measure $p(y|x)$ is by using ABX psychophysical tests. In an ABX test, listeners are presented with a sequence of three syllables, differing only in one phoneme. The first syllable contains the phoneme x_1; the second contains the phoneme x_2; the third contains either x_1 or x_2. The listener is asked to specify whether the third phoneme is x_1 or x_2. By guessing uniformly at random, a listener can be correct 50%

of the time; a native of the target language is typically correct about 99% of the time [9]. The Perceptual Assimilation Model proposes several broad categories of cross-lingual mappings, and experimental tests of the theory have demonstrated that the probability of error in an ABX task is much greater for some mapping categories than others. The types of mappings proposed in [8] include the two-category (TC) mapping, the non-assimilable (NA) mapping, the category-goodness (CG) mapping, and the one-category (OC) mapping.

A **two-category** (TC) mapping is one in which $y_1 \neq y_2$. An example of a TC mapping is the distinction between voiced versus unvoiced lateral fricatives in the Zulu language (/ɮ/ and /ɬ/), as perceived by native speakers of English. English does not have lateral fricatives, so both $x_1 =$/ɮ/ and $x_2 =$/ɬ/ are foreign. For most English-speaking listeners, however, the voiced Zulu lateral fricative sounds like $y_1 =$/z/ or $y_1 =$/ʒ/, while the unvoiced fricative sounds like $y_2 =$/s/ or $y_2 =$/ʃ/. Because $y_1 \neq y_2$, non-speakers of Zulu are very good at distinguishing these sounds, with an ABX accuracy of about 95% [8].

A **non-assimilable** (NA) mapping is one in which neither x_1 nor x_2 sounds like any phoneme in the listener's language; we can write $y_1 = \emptyset$, $y_2 = \emptyset$. An example are the dental and lateral clicks in Zulu, /|/ and /‖/. Neither are used in English speech, but both are used in English non-speech communication; /|/ is a tsk-ing sound that parents make toward mildly naughty children, while /‖/ is a giddy-up sound that a rider makes to her horse. As non-phonemes, these sounds are not quite as easily discriminated as the members of a TC mapping, but still, discrimination is not too hard; listeners typically achieve ABX accuracy around 81–99% on tasks like this.

A **category goodness** (CG) mapping is one in which both x_1 and x_2 map to the same English phoneme, $y_1 = y_2 = y$, but in which x_1 sounds like a good example of the category y, while x_2 sounds like a really bad example: $D(y, x_1) \ll D(y, x_2)$. An example is the Amharic distinction between voiceless aspirated stops (/pʰ,tʰ,kʰ/) and voiceless ejective stops (/p',t',k'/). To an English-speaking listener, both sound like the English phones /pʰ,tʰ,kʰ/, Amharic /pʰ,tʰ,kʰ/ are an audibly better match than the Amharic /p',t',k'/. ABX accuracy on this task, as measured by [8], was 88%.

A **single category** (SC) mapping is one in which both x_1 and x_2 map to the same English phoneme, $y_1 = y_2 = y$, with no audible difference in the quality of the mapping, $D(y, x_1) \approx D(y, x_2)$. An example is the Zulu distinction between bilabial plosive stops, /b/, and bilabial implosive stops, /!b/. In Zulu, replacing one of these sounds with another can change the meaning of a word (i.e., they are distinct phonemes), but English speakers usually find these two distinct to be almost indistinguishable from the English /b/. Accuracy in this ABX discrimination task was measured to be 65%, i.e., low, but still significantly better than chance [8].

B. A Learned Model of 2^{nd}-Langage Speech Perception

Suppose we have several minutes of speech audio that has been transcribed by both a native speaker of the target language (producing native phone sequences $X = [x_1, \ldots, X_T]$), and by a series of L different mismatched transcribers (producing mismatched transcripts $Y_L = [y_{1l}, \ldots, y_{Tl}]$. Let us assume that the substitution error probability $p(Y|X)$ is a finite-memory process, and can therefore be modeled by a weighted finite state transducer (WFST). Let $S_L = [s_{1l}, \ldots, s_{Tl}]$ be the states in this transducer that are traversed in order to transduce the input sequence X to the output sequence Y_l. In our training data, X and Y_l are observed, but S_l is hidden. The goal is to learn a WFST that represents $p(Y|X)$ with minimum expected probability of error.

A WFST is learned by randomly initializing its weights, then by finding a new set of weights that improves the likelihood of the training data. Define an arc to be any unique combination of initial state, final state, input symbol and output symbol, thus $a_t = \{s_{t-1}, s_t, x_t, y_y\}$. The goal of WFST learning is to estimate the input-conditional arc transition probabilities, $p(y_t, s_t|s_{t-1}, x_t)$. Given an initial value of these probabilities, the probability of a phoneme sequence X being represented by English orthographic sequence Y_l is

$$p(Y_l|X) = \sum_{S_l} \prod_{t=1}^{T} p(y_{tl}, s_{tl}|s_{(t-1)l}, x_t) \tag{4}$$

and the posterior probability of a particular arc transition, given the observations, is

$$\gamma_{tl}(s', s|X, Y_l) = p(s_{tl} = s', s_{(t-1),l} = s|X, Y_l) \tag{5}$$

The arc probabilities can then be re-estimated as

$$p(y, s'|s, x) = \frac{\sum_{l=1}^{L} \sum_{t=1}^{T} \gamma(s', s|X, Y_l) [\![y_{tl} = y, x_t = x]\!]}{\sum_{s'} \sum_{l=1}^{L} \sum_{t=1}^{T} \gamma(s', s|X, Y_l) [\![x_t = x]\!]} \tag{6}$$

where $[\![x_t = x]\!]$ and $[\![y_{tl} = y, x_t = x]\!]$ are unit indicator functions. An example of the single-phone substitution probabilities learned in this way was published in [7]. Context-dependent finite-state examples can be downloaded from https://github.com/uiuc-sst/PTgen.

III. Costs and Benefits of Mismatched Crowdsourcing

Table II shows the relative costs versus benefits of matched versus mismatched transcription, in six different languages. All phone error rates in this table have been previously published; estimates of the transcription cost have not been previously published. The first four languages (yue, hun, cmn, swh) were published in [5]. In that experiment, matched transcriptions were acquired from University of Illinois students, at a cost of $150 per hour of audio (for those who completed the task). Mismatched transcripts were acquired from Amazon Mechanical Turk, at a cost of $0.10 per ten-second transcription task. Mismatched transcription would therefore be much less expensive than matched transcription, except that

TABLE II
Lang=Language: yue=Cantonese, hun=Hungarian, cmn=Mandarin, swh=Swahili, amh=Amharic, din=Dinka. PER_0=phone error rate (PER) of the best cross-language baseline, PER_1=phone error rate of a system adapted using either matched or mismatched transcription, $\Delta\%= 100(\mathrm{PER}_0 - \mathrm{PER}_1)/\mathrm{PER}_1$. mm(hrs)=hours of mismatched transcription, m(hrs)=hours of matched transcription in the target language. \$/hr=transcription cost per hour of audio, \$/%=transcription cost per one-percent reduction in PER.

Lang	PER_0	PER_1	Δ%	mm(hrs)	m(hrs)	$/hr	$/%
yue	66.59	53.64	19	0.66	0	180	6.10
yue	66.59	27.67	58	0	0.66	150	1.69
hun	66.43	56.70	15	0.66	0	180	8.10
hun	66.43	35.87	46	0	0.66	150	2.15
cmn	65.77	54.07	18	0.66	0	180	6.70
cmn	65.77	27.80	58	0	0.66	150	1.71
swh	65.30	44.73	32	0.66	0	180	3.75
swh	65.30	34.98	46	0	0.66	150	2.13
amh	66.53	59.48	11	0.66	0	180	11.20
amh	66.53	43.92	34	0	0.66	870	16.90
din	64.78	58.22	10	0.66	0	180	11.75
din	64.78	48.58	25	0	0.66	700	18.47

TABLE III
Dollar costs and PER benefits of several different ASR systems developed for the Singapore Hokkien language. m(hrs)=hours of matched transcription, m(\$)=dollar cost of matched transcription, mm(hrs)=hours of mismatched transcription, mm(\$)=dollar cost of mismatched transcription, Total \$=total dollar cost of training data, PER=phone error rate.

m(hrs)	m($)	mm(hrs)	mm($)	Total $	PER
0.2	157.17	0	0.00	157.17	68.59
0.4	314.34	0	0.00	314.34	61.18
0.8	628.67	0	0.00	628.67	56.07
0.2	157.17	3.0	1035.42	1192.59	58.70
0.4	314.34	3.0	1035.42	1349.76	55.33
0.8	628.67	3.0	1035.42	1664.09	53.24
0.2	157.17	0.5	172.57	329.74	68.05
0.2	157.17	0.75	258.86	416.02	66.60
0.2	157.17	1	345.14	502.31	65.95
0.2	157.17	1.5	517.71	674.88	64.55

every segment of mismatched transcription was repeated by five different transcribers; the total cost for mismatched transcription is therefore \$0.10 × 360 × 5 =\$180 per audio hour. The ASR, in this experiment, is a hidden Markov model with 7800 clustered triphone states and a phone bigram language model. The acoustic model is a seven-layer feedforward neural network, 1024 nodes/layer. Acoustic features are log filterbank coefficients, stacked across five frames then transformed using linear discriminant analysis followed by a speaker-adaptive linear transform, then stacked again across five frames. Mismatched transcripts are time-aligned to form a probabilistic transcription, whose edges are trimmed to retain only those with probability greater than 0.7; edge probabilities then serve as soft targets for neural network training.

For the first four languages in Table II, therefore, mismatched transcription (with 5 transcribers per segment) looks more expensive than matched transcription (with 1 transcriber per segment). The apparent low expense of matched transcription, however, is an artifact: in this experiment, we started with audio in seventy languages, and developed ASR only in the 6 languages for which it was possible to hire a transcriber (at \$150/audio hour). In the other 64 languages of our corpus, it was not possible to hire a transcriber at a rate of \$150/audio hour.

The last two languages in Table II, Dinka=din and Amharic=amh, were systems developed for the paper [10]. This paper was written for a special session on African languages, therefore we advertised on upwork.com to find transcribers specifically capable of working in Amharic and Dinka. It was possible to hire transcribers in these languages, but only at a much higher rate: \$700/audio hour for the Dinka transcriber, \$870/audio hour for the Amharic transcriber. At these rates, mismatched transcription begins to look more cost-effective than matched transcription, even with a redundancy of 5 mismatched transcribers per audio segment. It is also possible that the improved cost-benefit ratio of mismatched transcription, in this paper, is caused by improved methods for using mismatched transcription to train a neural network: in this paper the probabilistic transcript and matched transcripts were used as targets in separate softmax layers.

The relative costs and benefits of matched versus mismatched transcription were measured in a different way in the manuscript [11]. In that paper, we developed methods (including mismatched bottleneck features and multi-task learning) capable of using varying amounts of both matched and mismatched transcription in the same target language. The best resulting phone error rates, for several different data cost levels, are listed in Table III. In this experiment, matched transcriptions were acquired by hiring a native speaker of the Singapore Hokkien language for one month, teaching her a writing system for that language, and then asking her to transcribe data; she was able to transcribe 3.67 hours of data in total. Mismatched transcripts were acquired from native speakers of Mandarin, via Upwork.com; transcribers were asked to "write down what you hear as if it were nonsense speech in Mandarin, using the best-matching syllables in the Mandarin Pinyin writing system." Two transcribers per audio clip were hired at \$172.57 per transcriber per audio hour.

Fig. 2 shows two scatter plots, each containing one point for each row in Table III. The scatter plot on the left shows PER as a function of the total dollar cost; the plot on the right shows $\ln(\mathrm{PER})$ as a function of total cost. As shown, both of these relationships are strikingly linear: dollar cost explains $R^2 = 56.57\%$ of variance in PER, and $R^2 = 57.29\%$ of variance in $\ln(\mathrm{PER})$.

The linear relationship is violated by only two points: the systems with more matched transcription (0.4 and 0.8 hours), and with no mismatched transcription, are able to achieve PER that is not quite as good as the best tested system, but that has a better cost-benefit tradeoff than the best tested system. In both cases, the data are much better fit by a pair of lines than by a single line ($R^2 \approx 70\%$ for the mismatched transcripts, $R^2 \approx 40\%$ for the matched transcripts in both cases).

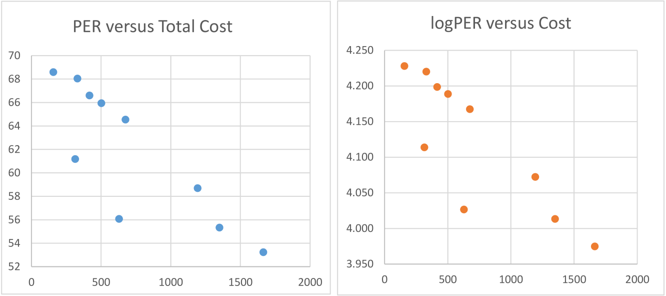


Fig. 2. Left: Phone error rate (PER) versus total dollar cost of available training data, ten different systems tested for Singapore Hokkien ($R^2 = 0.5657$). Right: log PER versus dollar cost ($R^2 = 0.5729$).

IV. A Cost-Benefit Model

Suppose that there are M different channels, each with cost/hour of C_m dollars: for example, the two channels might be mismatched transcription ($m = 0$) and matched transcription ($m = 1$). The scatter plots in Fig. 2 suggest two different possible cost-benefit models: either the error rate is a linear function of cost with a slope per channel of e_m error points per hour of data, or an exponential function of cost with a multiplicative factor of e_m percent relative reduction per hour of data. Suppose that we purchase f_m hours of data from the m^{th} channel. Then the two models are

$$\textbf{Model 1:}\quad E = E_0 - \sum_{m=1}^{M} e_m f_m \tag{7}$$

$$\textbf{Model 2:}\quad E = E_0 \exp\left(- \sum_{m=1}^{M} e_m f_m \right) \tag{8}$$

Given a total budget of C_{TOT}, the goal is to minimize E subject to a fixed total budget:

$$\{f_m^*\} = \arg\min E \quad \text{s.t. } C_{TOT} = \sum_{m=1}^{M} C_m f_m \tag{9}$$

Both models lead to the same Lagrangian:

$$\mathcal{L} = \sum_{m=1}^{M} e_m f_m - \lambda \sum_{m=1}^{M} C_m f_m \tag{10}$$

Optimizing Eq. 10 is a linear programming problem, whose optimum point is reached at one of the vertices of the feasible set:

$$f_m^* = \begin{cases} \frac{C_{TOT}}{C_m} & m = \text{argmin}\left(\frac{C_m}{e_m}\right) \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

V. Conclusions

Mismatched crowdsourcing is a proposed new method for acquiring speech transcriptions in an under-resourced language. In particular, when there are no transcribers available in the target language, or when the cost of transcription is very high, then we ask non-speakers of the language to write down what they hear as if the speech were nonsense speech in their own language. Misperceptions are interpreted on the basis of a model of second-language speech perception.

The relative utility of mismatched and matched crowdsourcing are measured by the error rate of the trained ASR. This paper has proposed two different models of the cost-benefit ratio, which seem to be identically useful: (1) phone error rate is a linear function of dollar cost, (2) phone error rate is an exponential function of dollar cost. In both cases, the slope of the curve relating error to cost may differ depending on whether the data come from matched versus mismatched crowdsourcing. Analysis of the cost-benefit ratio suggests that the most effective strategy is to find out which of the two transcription methods has the lowest price per incremental reduction in phone error rate, and to acquire all of our resources using this channel.

Utility of this analysis is limited, in practice, because it is hard to know in advance whether matched or mismatched transcription has the lowest cost per reduction. Table II showed examples of four languages (Hungarian, Mandarin, Cantonese and Swahili) for which matched transcription was most effective, and two languages (Amharic and Dinka) for which mismatched transcription was most effective.

Acknowledgment

The first author was supported during the development of this research by a grant from the DARPA LORELEI program.

References

[1] LE Baum and JA Eagon, "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology," *Bull. Am. Math. Soc.*, vol. 73, pp. 360–363, 1967.

[2] LR Rabiner, JG Wilpon, and B-H Juang, "A segmental k-means training procedure for connected word recognition," *Bell Labs Technical Journal*, vol. 65, no. 3, pp. 21–31, 1986.

[3] Hong Yu Qing Amelia, "A Phonological and phonetic Description of Singapore Hokkien," *B. A. thesis, Nanyang Technological University*, 2012.

[4] V Lim, HS Ang, E Lee, and B-P Lim, "Towards an Interactive Voice Agent for Singapore Hokkien," *HAI '16 Proceedings of the Fourth International Conference on Human Agent Interaction*, pp. 249–252, 2016.

[5] M Hasegawa-Johnson, P Jyothi, D McCloy, M Mirbagheri, G di Liberto, A Das, B Ekin, C Liu, V Manohar, H Tang, EC Lalor, N Chen, P Hager, T Kekona, R Sloan, and AKC Lee, "ASR for under-resourced languages from probabilistic transcription," *IEEE/ACM Trans. Audio, Speech and Language*, vol. 25, no. 1, pp. 46–59, 2017.

[6] M Hasegawa-Johnson, J Cole, P Jyothi, and LR Varshney, "Models of dataset size, question design, and cross-language speech perception for speech crowdsourcing applications," *Laboratory Phonology*, vol. 6, no. 3-4, pp. 381–432, 2015.

[7] P Jyothi and M Hasegawa-Johnson, "Acquiring speech transcriptions using mismatched crowdsourcing," *Proc. AAAI*, 2015.

[8] CT Best, GW McRoberts, and NN Sithole, "Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by english-speaking adults and infants," *J. Exp. Psychol. Human Percept. Perform.*, vol. 14, pp. 345–360, 1988.

[9] GA Miller and PE Nicely, "Analysis of perceptual confusions among some english consonants," *J. Acoust. Soc. Am.*, vol. 27, pp. 338–352, 1955.

[10] A Das, P Jyothi, and M Hasegawa-Johnson, "Automatic speech recognition using probabilistic transcriptions in Swahili, Amharic and Dinka," in *Proc. Interspeech*, 2016, pp. 3524–3527.

[11] VH Do, NF Chen, B-P Lim, and M Hasegawa-Johnson, "Multi-task learning for phone recognition of under-resourced languages using mismatched transcription," in review, 2017.