Language Coverage for Mismatched Crowdsourcing

Lav R. Varshney, Preethi Jyothi, and Mark Hasegawa-Johnson
Beckman Institute for Advanced Science and Technology
University of Illinois at Urbana-Champaign

***Abstract*—Developing automatic speech recognition technologies requires transcribed speech so as to learn the mapping from sound to text. It is traditionally assumed that transcribers need to be native speakers of the language being transcribed. Mismatched crowdsourcing is the transcription of speech by crowd workers who do not speak the language. Given there are phonological similarities among different human languages, mismatched crowdsourcing does provide noisy data that can be aggregated to yield reliable labels. Here we discuss phonological properties of different languages in a coding-theoretic framework, and how nonnative phoneme misperception can be modeled as a noisy communication channel. We show the results of experiments demonstrating the efficacy of this information theory inspired modeling approach, having native English speakers and native Mandarin speakers transcribe Cantonese speech. Finally we discuss how crowd workers whose native language background give them the highest probability of faithful transcription can be found by solving a weighted set cover problem.**

***Index Terms*—channel selection, distance distribution, phonology, mismatched crowdsourcing, set cover, speech transcription**

I. Introduction

There are more than six billion people in the world speaking more than six thousand living languages, but very few languages are spoken by more than a few thousand people each [1]. Speech technology in minority languages has the potential to empower minority communities, and in some cases promote language preservation and diversity. In particular, speech is an ideal medium for human-machine interaction since it is very natural for people, whether literate or illiterate. Developing speech technology, however, is data intensive.

To develop speech technology, it is necessary to have some amount of recorded speech audio, some amount of text written in the target language, and a *transcription* mapping that links sound to text. Although speech audio can be recorded during weekly minority-language broadcasts on local radio stations, and text can be acquired from printed pamphlets and literacy primers, the link is usually missing. Moreover, native language transcription is beyond the economic capabilities of many minority-language communities.

Paucity of labeled training data is problematic not just for speech, but also for many other machine learning problems. Recent demonstrations of deep learning have captured the public imagination by approaching human-level performance in image classification, but what is lost in the magic of TED talks and social media company demonstrations is the exceedingly large amount of human effort that goes into creating labeled training data. In order to run the ImageNet Large-Scale Visual Recognition Challenge, human workers from Amazon Mechanical Turk were recruited to classify 1.2 million training images, 50 thousand validation images, and 100 thousand test images into 1000 object classes reliably despite their own individual unreliability [2].

So much effort for a single general task that any person is qualified to do; similar effort may be needed for each different machine learning task that arises, such as labeling phonemes in thousands of languages where speech technologies are desired. We could draw on a crowdsourcing platform, but there is a significant *mismatch* between the number of native speakers of languages in the world and the number of crowd workers who speak that language, see Fig. 1. Moreover, native speakers are often more willing to record a language than they are to transcribe it, especially in the quantities necessary for modern speech technology. The cognitive surplus in the world has a distributional expertise mismatch with the expertise required to complete tasks. How should we proceed?

Although there is growing interest in transfer learning [4], we have developed a method that altogether bypasses the need for native language transcription. Our method, which we call *mismatched crowdsourcing*, proposes that speech in a target language be transcribed by crowd workers who have no knowledge of the target language, and that explicit models of second-language perception be used to recover an equivalent transcription in the language of the speaker [5], [6]. The reason this may work is that phonemes are not abstract symbols but have attributes along various phonological dimensions, and many languages share these phonological dimensions. Thus, even people that do not speak a language can correctly perceive certain phoneme attributes.

Explicit decomposition of labeling tasks into subtasks based on parts and attributes has previously been developed as a way to mitigate lack of expertise in the crowdsourcing of fine-grained tasks [7]. We have also made use of the factoring of tasks into easier subtasks to develop error-correcting codes for crowdsourcing [8].

Here the workers are not given an explicit decomposition of the task into factors; rather we study properties of the decomposition for speech transcription to determine whether aggregation techniques can implicitly make use of attributes. To do so, we use a coding-theoretic framework to study the phonological structure of several world languages, and to determine how nonnative phoneme misperception can be modeled as a noisy communication channel.

To validate our theoretical development, we also conduct

This work was supported in part by NSF Grant IIS-1550145.

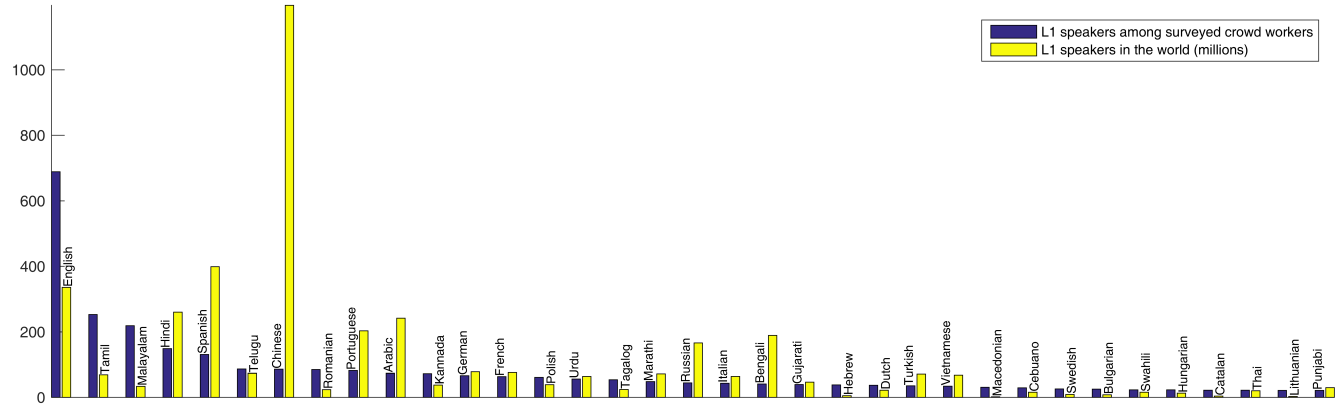

Fig. 1. There is a distributional mismatch in native language between crowd workers surveyed on Amazon Mechanical Turk [3, Table 1] and the native language populations around the world [1]. Native languages with more than 20 speakers among crowd workers in Pavlick, et al.'s survey are shown. Note, e.g., that there are proportionally more workers than population in English, Tamil, and Malayalam, whereas there are less in Hindi, Spanish, and Chinese. Taking numbers for these 35 languages as probability mass functions, the $\log_2$-relative entropy between the world at large and the crowd workforce is 1.0846.

experiments with human subjects (we have obtained clearance from the Institutional Review Board for the Protection of Human Subjects at the University of Illinois at Urbana-Champaign). We demonstrate there is indeed differential performance in mismatched transcription, depending on the language background of workers. In particular, performance is partly explained by the absence/presence of discriminative phonological dimensions in the workers' native language. The experiment involves English and Mandarin native workers transcribing Cantonese.

Finally, we use our framework of phonological structure to determine how to find crowd workers whose native language background give them the highest chance of faithful transcription, by casting as a weighted set cover problem, which also takes into account the native language distribution of workers in the crowd. In addition to Amazon Mechanical Turk (which is common in speech applications [9], [10]), we also consider UpWork (formerly oDesk) which is a freelance market [11] that allows workers to list skills. Note that although there is individual variation among people [12], we characterize workers simply by their native language.

A. Language and Information Theory

This work continues a long strand of research at the intersection of linguistics and information theory. Shannon developed stochastic language models via human-based prediction and used these models to estimate the entropy of English [13]; the same human-based prediction approach—Shannon's guessing game—has been used to estimate entropy of other languages such as Hindi [14]. In one study reminiscent of mismatched crowdsourcing, the person playing Shannon's guessing game does not know the language to be predicted, only knowing their native language. Performance is said to indicate the level of similarity between the native language of the guesser and the language being guessed [15]. The statistical methodology of Shannon's guessing game can be improved using gambling rather than guessing [16]. Note that this stochastic approach has classically been criticized by Chomsky as not descriptive of human language [17].

All of these studies were concerned with orthographic symbols, rather than sounds. There has been study of phonology using the tools of statistical communication theory, dating back to Cherry, Halle, and Jakobson [18] where ideas of phonological dimensions were discussed. However, this line of investigation was not pursued further in the communication theory literature. The idea that some languages are closer to others also has a long history and is motivated by disparate concerns such as in historical linguistics [19] or in international trade and migration [20]. Notions of language similarity have been particularized to phonetic similarity [21]; we will revisit this in an information-theoretic framework as we pursue our engineering concern of building automatic speech recognition technology for under-resourced languages.

II. Distinctive Features and Coding Theory

A. Background

In phonetics and phonology, dating back to the theoretical framework of the ancient Sanskrit grammarian Pāṇini, spoken language is represented as abstract segments that are discrete and serially ordered. Segments are phonological representations that consist of distinctive features [22]–[24]; they are abstractions of articulatory or auditory units of speech production or perception. Segmental phonology studies the distribution of speech sounds and their patterning to understand how contrastive sounds trigger lexical or grammatical differences in languages. A key theoretical construct is the *phoneme*, a minimally distinctive sound in a particular language having a set of contrastive segments based on phonological principles. Each spoken language uses a set of consonants and vowels to form words; this set is called a *segment inventory*. A segment inventory implicitly encodes the phonetic dimensions employed by a phonological system to form meaningful contrasts.

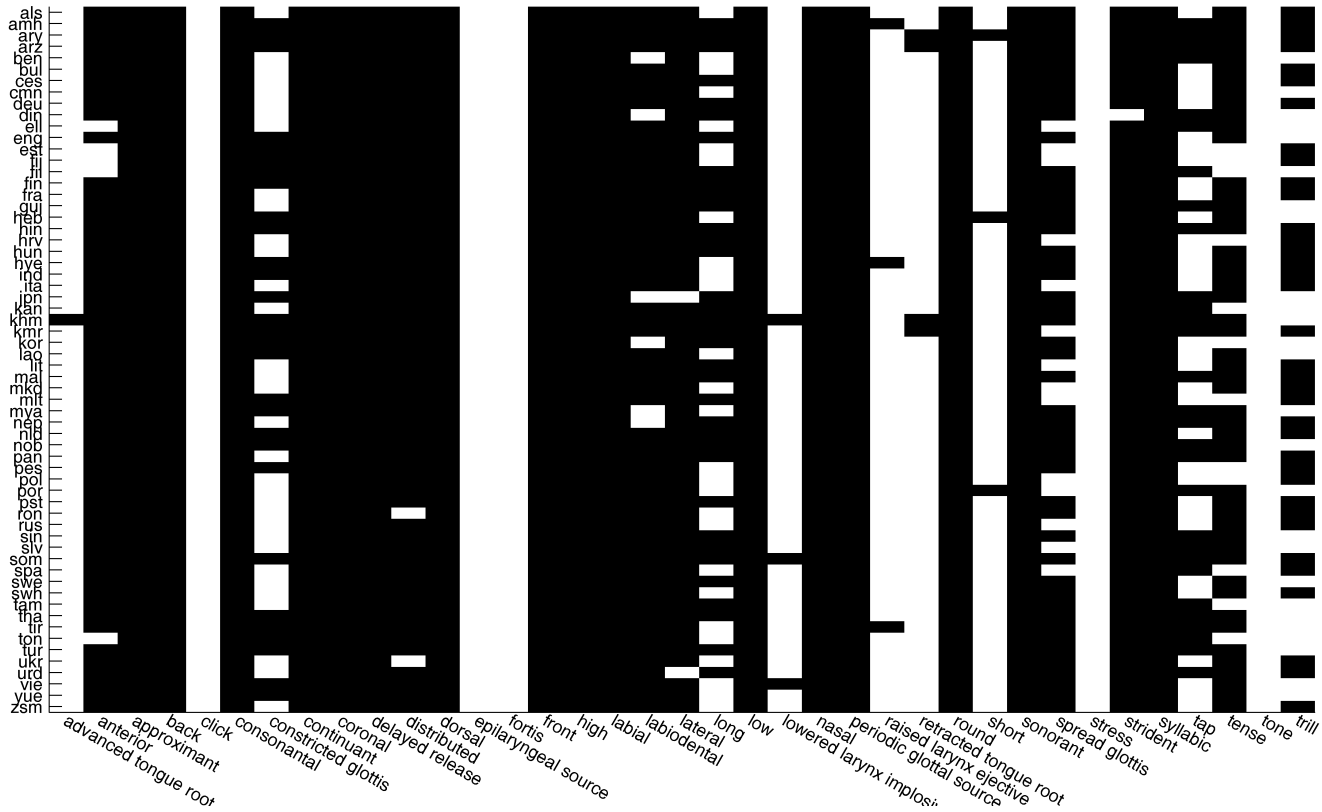

Fig. 2. Contrastive phonological features in several languages. Languages are represented by their ISO 639-3 codes, given in Table I.

Distinctive features represent abstract properties of speech sounds, modeled as binary feature values. Features are viewed as anatomically grounded in that they correspond to articulatory settings that have relatively stable, distinctive acoustic properties. As segments are bundles of distinctive features, two speech sounds contrast if they differ by at least one distinctive feature. The *feature matrix* expresses contrasts among speech sounds by distinctive features; the matrix can be used to calculate how much two segments differ by summing up the oppositions of their features, i.e. Hamming distance. That is, this is equivalent to a code matrix in coding theory.

The International Phonetic Alphabet has symbols as shorthand for representing articulatory features, but as an abstract set of symbols. Featural writing systems are less common and encode distinctive features within the shapes of symbols in the script; Korean Hangul is a prominent example.

Cross-linguistic comparisons of segment inventories provide insights into the factors that shape phonology in human language. For example, it has been well-noted that the set of consonants and vowels that can make up a segment inventory are constrained [25] and occur with varying frequencies in languages around the world [26]. Recently a comprehensive database of cross-linguistic phonological inventory data, Phonetics Information Base and Lexicon (PHOIBLE), has been compiled from source documents and tertiary databases [27]. The 2014 edition of PHOIBLE, of which we use a subset in the sequel, includes 2155 inventories that contain 2160 segment types found in 1672 languages, as well as distinctive feature data for every phoneme in every language [28]. We have drawn Fig. 2 to illustrate which distinctive features are used in several global languages of varying popularity.

B. Noisy Channel Model

Following distinctive feature theory, let us think of communication as binary symbols along phonological dimensions perturbed by a noisy channel, rather than full phonemes going through a noisy channel. The ability to perceive foreign phonological contrast dimensions is largely lost after the age of one year [29]–[31]. Hence, we can model a mismatched worker as a low-noise binary symmetric channel for phonological dimensions that are discriminative in the native language, and as a very noisy binary symmetric channel for phonological dimensions that are not discriminative in the native language. In the extreme, this would be pure noise for foreign dimensions and noiseless communication for native dimensions. In a sense, this gives erasures for nonnative dimensions (but without a special erasure symbol), and it is known that maximum distance separable (MDS) codes are optimal for the binary erasure channel.

Let us consider an example. Of the 27 phonological dimensions in Hindi, 25 are also used in English (tap and trill are not used), so roughly 2 symbols of information may be erased in transmission. Since there may be redundancy in the transmission procedure, less than 2 information bits may

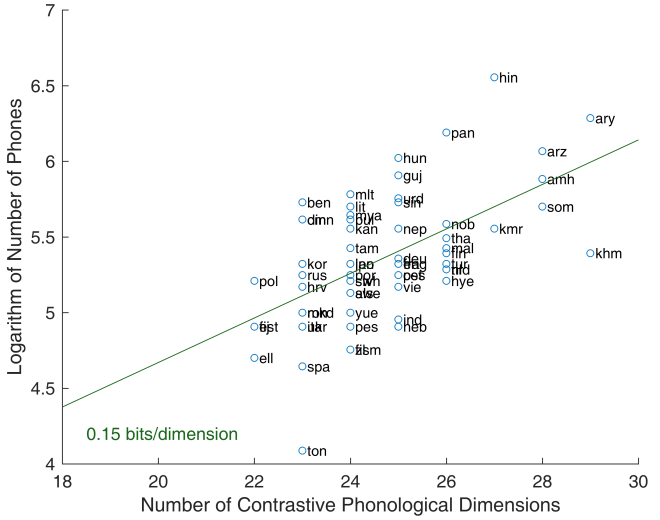

Fig. 3. Entropic characterization of the phonology in several languages. Languages are represented by their ISO 639-3 codes, given in Table I.

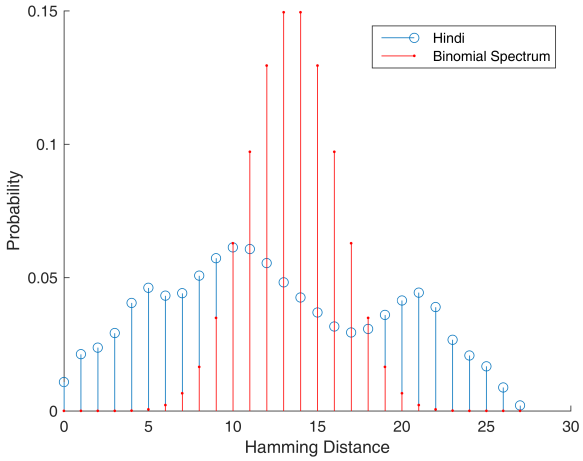

Fig. 4. Probability mass function (cdf) of the Hamming distance (in terms of phonological dimensions) for Hindi; some of the bimodal nature is explained by differences between vowels and consonants. A binomial spectrum for length 27 and 94 codewords is shown for comparison; clearly Hindi is not a random code.

actually be erased. Nevertheless, we can make a comparison to our past experiments in mismatched crowdsourcing where Hindi was transcribed by native English speakers that did not know any other language [6]. We saw that equivocation (conditional entropy) of English letters given Hindi phones for different phone classes in experimental data is 2.90 bits, so erasures of phonological dimensions do explain much of the loss. After all, even native Hindi transcribers from the crowd are unreliable and would leave some equivocation. This remaining equivocation should be mitigated by a second worker who corrects the result [32, Fig. 8], whether a Hindi speaker or another mismatched worker that covers the remaining uncertainty (as we will see in Sec. IV).

We can also note that there may be some loss due to the mismatched crowd worker mapping from sounds into his own native orthography, constructed from *graphemes* that are the basic, minimally distinctive symbol of a particular writing system. There are models of phoneme-to-grapheme transduction that may be used for correcting such errors [6].

C. Explanatory Principles

It is clear that if there are n binary distinctive features, there can be a maximum of 2^n distinctive sounds in a language; this is called the *feature bounding principle* in phonology [33]. Thinking of the segment inventory of a language as a binary code in Hamming space H_2^n, one might wonder if it densely fills the space, in terms of the level of redundancy.

The *feature economy* principle in phonology suggests that there should not be too much redundancy in the code, as quantified by a measure called the economy, the ratio of the number of sounds to the number of features [34]. To measure feature economy, we consider its logarithm for several languages in Fig. 3. Note that although measured in bits, Fig. 3 does not take frequency of phoneme occurrence into account, and so is an upper bound on entropy. Phoneme frequencies in several languages seem to follow Yule distributions [35], and so the entropy is much less than the logarithm of the number of phonemes (see [36] for a phoneme entropy computation for Hindi). As we observe, the inventories of languages are nowhere near 1 bit/dimension and there is a great deal of redundancy, so the feature economy principle alone is insufficient as an explanatory principle.

In error-correcting codes, redundancy is used to reduce decoding error, and so perhaps this is also the case in languages. This idea, having well-differentiated phonetic dimensions so that members of an inventory are highly individualized and distinct from one another, is called the *robustness principle* in phonology [33]. In fact, distance has been proposed as an

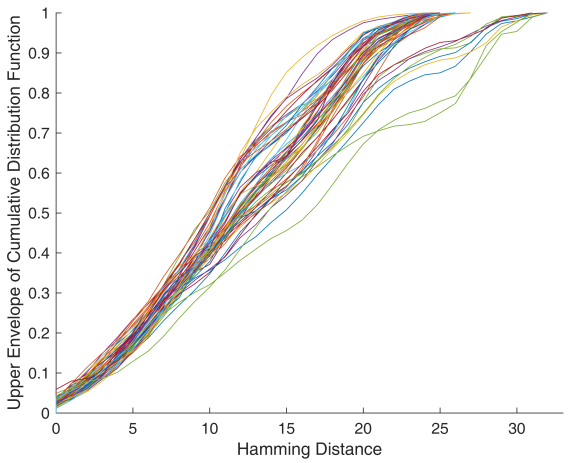

Fig. 5. Cumulative distribution function (cdf) of the Hamming distance (in terms of phonological dimensions) for several languages (all listed in Table I). So it is easier to see the distinct functions which comprise staircase steps, the upper envelope for each cdf is shown.

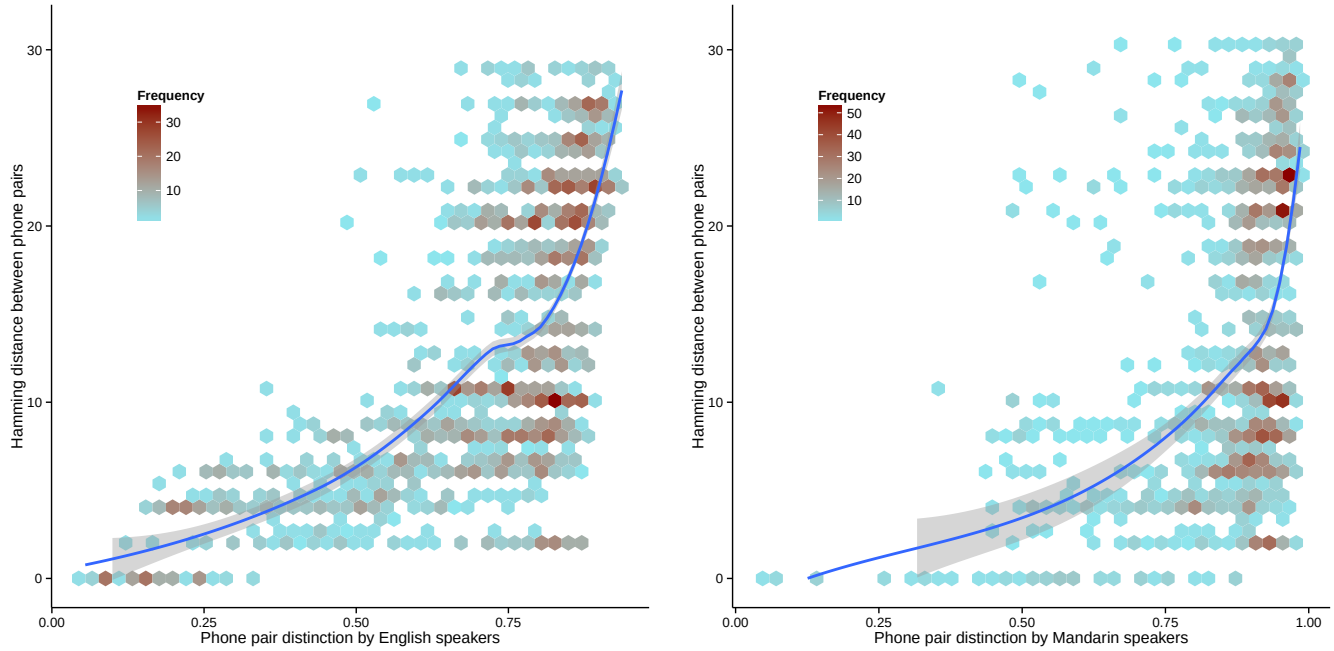

Fig. 6. Phone pair distinction plotted against Hamming distance of phonological feature vectors restricted by transcribers' native language. Data derived from Cantonese speech transcribed by English speakers (on the left) and Mandarin speakers (on the right).

explicit optimization criterion for phonology [37], just like in algebraic coding theory. If languages were optimal in the sense of coding theory—achieving the Singleton bound for binary codes—a language like Hindi with 94 codewords and length $n = 27$ would have a minimum Hamming distance of $27 + 1 - \log_2(94) \approx 21$, but as we will see in Figs. 4 and 5, this is not the case either. So the robustness principle alone does not explain language phonology.

It has also been observed that segment inventories in languages are spread out in feature space [38], so one might wonder if inventories are just random codes. The average distance distribution of a code chosen in the Hamming space H_2^n with uniform probability is the binomial spectrum, $A_w = \binom{n}{w}|C|/2^n$, as noted e.g. in [39]. As we see in Fig. 4, Hindi phonology is clearly not random. There is structure.

Minimal redundancy, maximal minimum distance, and random coding do not explain observed phonologies, but if we look at the distance distribution of several languages in Fig. 5, it appears there is some sort of universal law that governs the phenomenon. An open question is to find an explanatory theoretical framework. Previous proposals to explain spreadness of inventories include the idea that phonemes are like particles that repel each other in space [40], similar to energy-minimizing error-control codes [41]; and a successive division algorithm operating in continuous feature space [38], similar to tree-structured vector quantization [42].

III. Effect of Native Language on Nonnative Phone Differentiation

In this section, we investigate the misperception of phones by mismatched transcribers. Towards this, we devise an empirical *phone pair distinction* measure, $\delta(\alpha, \beta)$, between a pair of phones $\{\alpha, \beta\}$. We explore whether $\delta(\alpha, \beta)$ is correlated with the distance between α and β in the phonological feature space. We also use this measure to visualize the effect of the native language background of crowd workers on the misperception of foreign sounds.

For our experiments, we choose Cantonese to be our foreign language and we employ two sets of crowd workers with different language backgrounds: English speakers (employed on Amazon Mechanical Turk) and Mandarin speakers (employed on UpWork). Each crowd worker listens to a short speech clip in Cantonese and provides a transcription that is acoustically closest to what they think they heard. The transcriptions from the English speakers are in English (mostly in the form of nonsense syllables and not corresponding to valid English words) and the Mandarin speakers use the Pinyin alphabet. Using these two independent sets of transcriptions, we estimate separate noisy channel models of Cantonese phoneme misperception [6].

A Cantonese phone sequence can be aligned with its corresponding mismatched transcription (either in English or Pinyin) using these trained channel models. For a pair of Cantonese phones $\{\alpha, \beta\}$, $\delta(\alpha, \beta)$ is defined as the total variation distance between probability distributions over symbols that are assigned to the two phones in the aligned transcriptions. Let S_α denote the probability distribution aligned to α, over alphabet $\mathcal{S}$; similarly, let S_β be the symbol distribution corresponding to β. The total variation distance, $\Delta(S_\alpha, S_\beta) = \frac{1}{2}\sum_{x\in\mathcal{S}} |S_\alpha(x) - S_\beta(x)|$. Then $\delta(\alpha, \beta) = 1$ when the phones α and β can be perfectly discriminated and $\delta(\alpha, \beta) = 0$ when α and β have identical symbol distributions.

We define distance between two phones $\{\alpha, \beta\}$ in the phonological space, $\eta(\alpha, \beta)$, as the Hamming distance between their feature vectors. We restrict the feature vectors to those phonological dimensions that appear in the transcribers' native language. To visualize the correlation (if any) between $\delta(\alpha, \beta)$ and $\eta(\alpha, \beta)$, we plot each Cantonese phone pair $\{\alpha, \beta\}$ as a point $(x, y) = (\delta(\alpha, \beta), \eta(\alpha, \beta))$. Fig. 6 shows a heat map where regions with a higher concentration of such points are shown in darker colors. The plots also include a regression line fit to the data.[1]

As is evident from Fig. 6, there is a clear positive correlation between $\delta(\alpha, \beta)$ and $\eta(\alpha, \beta)$. This correlation holds for transcriptions by both the English speaking and Mandarin speaking transcribers. This suggests that it is reasonable to model the channel in terms of the phonological features of the phones being transmitted. Comparing the two plots in Fig. 6, we see a similar trend across both sets of transcribers despite the experiments being conducted in different settings.[2] However, we also observe that points in the plot for Mandarin speakers are shifted to the right. This suggests that, on average, the Mandarin-speaking transcribers can distinguish between a pair of Cantonese phones better than the English-speaking transcribers, even if the distance between the phones is similar in the respective phonological feature spaces. This indicates that the channel model is not solely characterized by the unweighted Hamming distance between phonological feature vectors. Gaining full understanding is an open problem for future investigation.

A direct comparison between the phone-pair distinction by the two sets of transcribers is given in Fig. 7. The horizontal and vertical axes corresponds to $\delta(\alpha, \beta)$ computed using the transcripts from, respectively, the English-speaking transcribers and the Mandarin-speaking transcribers. The warmer colors in the heat map indicate cells containing larger number of phone pairs. We observe in Fig. 7 that most phone pairs are well-differentiated by both sets of transcribers. We also observe that there are many phone pairs which are not well-differentiated by English transcribers, but are significantly differentiated by Mandarin transcribers (i.e., have a low δ value on the horizontal axis and a high δ value on the vertical axis).

IV. Choosing Transcribers

Sec. II-B had developed a simple erasure channel model for mismatched crowdsourcing based on native/nonnative phonological dimensions, and Sec. III experimentally demonstrated the reasonableness of the model. Based on the simple model, we now develop an approach to select mismatched crowd workers that would be most effective for a given source language. We also take the potential scarcity of certain kinds

[1]The regression line was generated using a local polynomial regression fit algorithm implemented in R by the function `loess`.

[2]The English transcriptions from the English speakers on Amazon Mechanical Turk were derived for 5-second Cantonese clips by first splitting the clip into four roughly equal-sized splits and concatenating their respective mismatched transcriptions. On the other hand, the Pinyin transcriptions from the Mandarin speakers on UpWork were acquired without splitting the clips.

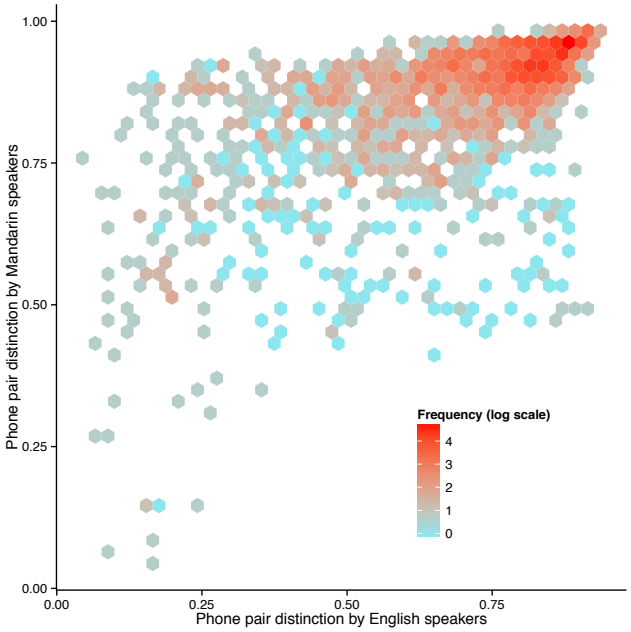

Fig. 7. Phone pair distinction in Cantonese using both English and Pinyin transcriptions.

of workers into account as weights in the optimization. The problem formulation is similar to channel selection in wireless communication systems, but due to our erasure model, it is combinatorial in nature.

The selection of transcriber languages for a given source language can be viewed as a weighted set cover problem: Select the minimal-cost set of transcriber languages from a universe of languages $\mathcal{L}$ such that all the phonological features of a source language L_0 are covered. Formally, this can be written as:

$$S^* = \min_{S \subseteq \mathcal{L}} \sum_{L \in S} w_L \quad \text{s.t.} \quad \bigcup_{L \in S} \Phi_L \supseteq \Phi_{L_0} \tag{1}$$

where w_L is the cost associated with the language L and Φ_L is the set of phonological features corresponding to L. The set S^* corresponds to the best set of languages selected to cover all the phonological features corresponding to L_0.

Although the decision version of set cover is NP-complete and the optimization version is NP-hard, our optimization problem can be set up as an integer linear program and solved exactly.[3] Table II shows the results of solving (1) for all languages listed in Table I. We consider an unweighted and weighted version of the problem. In the unweighted version, for each source language L_0, we let $\mathcal{L}$ be the set of all languages *excluding* L_0, and the costs w_L to be uniformly 1. In the weighted version, consider $\mathcal{L}$ to be the entire list of languages, and the costs w_L to be inversely proportional to the number of translators available on UpWork for L (see Table I). In the weighted version of the problem, it is

[3]We use the integer linear programming solver implemented in Matlab by the function `intlinprog`.

TABLE I
LANGUAGES IN CORPUS

Language	Language (ISO 639-3)	Number of Translators
Albanian	als	374
Amharic	amh	99
Arabic (Moroccan)	ary	13518
Arabic (Egyptian)	arz	13518
Bengali	ben	180
Bulgarian	bul	1324
Czech	ces	0
Chinese (Mandarin)	cmn	3746
German	deu	0
Dinka	din	1
Greek	ell	1854
English	eng	159683
Estonian	est	0
Fijian	fij	0
Filipino	fil	5696
Finnish	fin	0
French	frav	27666
Gujarati	guj	0
Hebrew	heb	845
Hindi	hin	3647
Croatian	hrv	2083
Hungarian	hun	1286
Armenian	hye	0
Indonesian	ind	2827
Italian	ita	0
Japanese	jpn	5914
Kannada	kan	222
Khmer	khm	0
Kurdish	kmr	38
Korean	kor	0
Lao	lao	0
Lithuanian	lit	0
Malayalam	mal	0
Macedonian	mkd	0
Maltese	mlt	0
Burmese	mya	48
Nepali	nep	80
Dutch	nld	3717
Norwegian	nob	0
Punjabi	pan	0
Persian	pes	237
Polish	pol	0
Portuguese	por	8559
Pashto	pst	106
Romanian	ron	2472
Russian	rus	12366
Sinhalese	sin	32
Slovenian	slv	0
Somali	som	63
Spanish	spa	33043
Swedish	swe	0
Swahili	swh	751
Tamil	tam	797
Thai	tha	1320
Tigrinya	tir	8
Tongan	ton	0
Turkish	tur	2608
Ukrainian	ukr	0
Urdu	urd	2374
Vietnamese	vie	1470
Cantonese	yue	663
Malaysian	zsm	0

TABLE II
RESULTS OF UNWEIGHTED AND WEIGHTED SET COVER

Target	Unweighted Cover	Weighted Cover
als	pst	fra
amh	pst, tir	amh
ary	heb, kmr	ary
arz	ary	arz
ben	tir	eng, spa
bul	pst	fra
ces	pst	fra
cmn	pst	eng
deu	pst	fra
din	hin	eng, spa
ell	tir	eng, spa
eng	hin	eng
est	hin	eng, spa
fij	hin	eng, spa
fil	hin	eng, spa
fin	hin	eng, spa
fra	pst	fra
guj	hin	eng, spa
heb	ary	ary
hin	som	eng, spa
hrv	pst	fra
hun	pst	fra
hye	amh	amh
ind	hin	eng, spa
ita	pst	fra
jpn	hin	eng, spa
kan	hin	eng, spa
khm	—	khm
kmr	ary	arz
kor	hin	eng
lao	hin	eng
lit	pst	fra
mal	hin	eng, spa
mkd	pst	fra
mlt	hin	eng, spa
mya	hin	eng, spa
nep	hin	eng, spa
nld	hin	eng, spa
nob	hin	eng, spa
pan	hin	eng, spa
pes	hin	eng, spa
pol	pst	spa
por	ary	ary
pst	hin	fra
ron	hin	fra
rus	hin	fra
sin	hin	eng, spa
slv	hin	eng, spa
som	hin, khm	eng, spa, vie
spa	hin	spa
swe	hin	eng
swh	hin	fra
tam	hin	eng, spa
tha	hin	eng, spa
tir	amh	amh
ton	hin	eng, spa
tur	hin	eng, spa
ukr	hin	fra
urd	hin	eng, spa
vie	som	vie
yue	hin	eng
zsm	hin	fra

best to use native speakers for languages with sufficiently many workers available—English, Spanish, French, Arabic, and Vietnamese—but not for any others. Khmer has an unusual phonological feature (advanced tongue root) that can only be covered by Khmer speakers in our worker pool.

This basic combinatorial channel selection problem can be extended to the setting where several languages need to be transcribed by several workers, so multiple covers are needed simultaneously [43].

V. Conclusion

In this paper, we have considered mismatched crowdsourcing of speech transcription as a noisy channel. Rather than thinking of phonemes as the alphabet that is perturbed by noise in the transcription process, we have considered binary symbols drawn from distinctive phonological dimensions, such that each phoneme corresponds to more than twenty binary symbols. Phonological dimensions that are native to the transcriber should have less noise than phonological dimensions that are not native. This modeling approach is borne out in experiments we have conducted with transcription of Cantonese by native English and native Mandarin speakers. Taking the model to the extreme of noiseless native dimensions and completely noisy nonnative dimensions yields a novel combinatorial channel selection problem that can be solved using an integer linear programming formulation of weighted set cover, which we apply to obtain an effective method to allocate specialized workers to tasks.

As part of our study, we also uncover basic open questions in understanding phonology. Distance distributions are used in coding theory to estimate error probabilities in decoding. We have looked at distance distributions of several world languages and there seems to be some universal non-random description that is unexplained by optimality principles in channel coding theory. For a principled explanation, perhaps the difficulty of production, cf. [44], must be considered. Perhaps a joint source-channel coding framework would provide more insight, whether taking the statistics of language sources into account, or through a combinatorial optimization [45]. We also note that Hamming distance between the feature vectors of phonemes does not fully explain the performance of human transcribers. A second open question is whether there is something more than unweighted distance that can be considered in channel modeling.

Acknowledgment

Thanks to Wenda Chen for providing Cantonese-Mandarin transcription data, worker pool data, and for discussions.

References

[1] G. F. Simons, M. P. Lewis, and C. D. Fennig, Eds., *Ethnologue: Languages of the World*, 18th ed. SIL International, 2015.

[2] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," arXiv:1409.0575 [cs.CV]., Sep. 2014.

[3] E. Pavlick, M. Post, A. Irvine, D. Kachaev, and C. Callison-Burch, "The language demographics of Amazon Mechanical Turk," *Trans. Assoc. Comput. Linguist.*, vol. 2, pp. 79–92, 2014.

[4] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[5] M. Hasegawa-Johnson, J. Cole, P. Jyothi, and L. R. Varshney, "Models of dataset size, question design, and cross-language speech perception for speech crowdsourcing applications," *Lab. Phonol.*, vol. 6, no. 3-4, pp. 381–431, Oct. 2015.

[6] P. Jyothi and M. Hasegawa-Johnson, "Acquiring speech transcriptions using mismatched crowdsourcing," in *Proc. 29th AAAI Conf. Artif. Intell.*, Jan. 2015, pp. 1263–1269.

[7] S. Branson, G. Van Horn, C. Wah, P. Perona, and S. Belongie, "The ignorant led by the blind: A hybrid human-machine vision system for fine-grained categorization," *Int. J. Comput. Vis.*, vol. 108, no. 1-2, pp. 3–29, May 2014.

[8] A. Vempaty, L. R. Varshney, and P. K. Varshney, "Reliable crowdsourcing for multi-class labeling using coding theory," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 4, pp. 667–679, Aug. 2014.

[9] M. Marge, S. Banerjee, and A. I. Rudnicky, "Using the Amazon Mechanical Turk for transcription of spoken language," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2010)*, Mar. 2010, pp. 5270–5273.

[10] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks," in *Proc. Conf. Empirical Meth. Natural Language Process. (EMNLP'08)*, Oct. 2008, pp. 254–263.

[11] A. Chatterjee, L. R. Varshney, and S. Vishwananth, "Work capacity of freelance markets: Fundamental limits and decentralized schemes," in *Proc. 2015 IEEE INFOCOM*, Apr. 2015, pp. 1769–1777.

[12] C. J. Fillmore, "On fluency," in *Individual Differences in Language Ability and Language Behavior*, C. J. Fillmore, D. Kempler, and W. S.-Y. Wang, Eds. New York: Academic Press, 1979, pp. 85–101.

[13] C. E. Shannon, "Prediction and entropy of printed English," *Bell Syst. Tech. J.*, vol. 30, no. 1, pp. 50–64, Jan. 1951.

[14] P. K. Varshney and A. Varshney, "On an information theoretic study of Hindi," in *Fourth South Asian Languages Roundtable*, May 1982.

[15] D. Jamison and K. Jamison, "A note on the entropy of partial-known languages," *Inf. Control*, vol. 12, no. 2, pp. 164–167, Feb. 1968.

[16] T. M. Cover and R. C. King, "A convergent gambling estimate of the entropy of English," *IEEE Trans. Inf. Theory*, vol. IT-24, no. 4, pp. 413–421, Jul. 1978.

[17] N. Chomsky, "Three models for the description of language," *IRE Trans. Inf. Theory*, vol. IT-2, no. 3, pp. 113–124, Sep. 1956.

[18] E. C. Cherry, M. Halle, and R. Jakobson, "Toward the logical description of languages in their phonemic aspect," *Language*, vol. 29, no. 1, pp. 34–46, Jan.-Mar. 1953.

[19] R. Georgi, F. Xia, and W. Lewis, "Comparing language similarity across genetic and typologically-based groupings," in *Proc. 23rd Int. Conf. Comput. Linguist. (COLING '10)*, Aug. 2010, pp. 385–393.

[20] I. E. Isphording and S. Otten, "The costs of Babylon—linguistic distance in applied economics," *Rev. Int. Econ.*, vol. 21, no. 2, pp. 354–369, May 2013.

[21] B. Kessler, "Phonetic comparison algorithms," *Trans. Philol. Soc.*, vol. 103, no. 2, pp. 243–260, Aug. 2005.

[22] R. Jakobson, C. G. M. Fant, and M. Halle, *Preliminaries to Speech Analysis: The Distinctive Features and their Correlates*. Cambridge, MA: MIT Press, 1952.

[23] R. Jakobson and M. Halle, *Fundamentals of Langauge*. The Hague: Mouton & Co., 1956.

[24] N. Chomsky and M. Halle, *The Sound Pattern of English*. New York: Harper & Row, 1968.

[25] E. Sapir, "Sound patterns in language," *Language*, vol. 1, no. 2, pp. 37–51, Jun. 1925.

[26] I. Maddieson, *Patterns of Sounds*. Cambridge: Cambridge University Press, 1984.

[27] S. P. Moran, "Phonetics information base and lexicon," Ph.D. dissertation, University of Washington, 2012.

[28] S. Moran, D. McCloy, and R. Wright, Eds., *PHOIBLE Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology, 2014.

[29] L. Polka and J. F. Werker, "Developmental changes in perception of nonnative vowel contrasts," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 20, no. 2, pp. 421–435, Apr. 1994.

[30] J. F. Werker and R. C. Tees, "Cross-language speech perception: Evidence for perceptual reorganization during the first year of life," *Infant Behav. Dev.*, vol. 7, no. 1, pp. 49–63, Jan.-Mar. 1984.

[31] C. T. Best, G. W. McRoberts, and E. Goodell, "Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system," *J. Acoust. Soc. Am.*, vol. 109, no. 2, pp. 775–794, Feb. 2001.
[32] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, July/Oct. 1948.
[33] G. N. Clements, "The role of features in phonological inventories," in *Contemporary Views on Architecture and Representations in Phonology*, E. Raimy and C. E. Cairns, Eds. Cambridge, MA: MIT Press, 2009, pp. 19–68.
[34] ——, "Feature economy in sound systems," *Phonology*, vol. 20, no. 3, pp. 287–333, Dec. 2003.
[35] C. Martindale, S. M. Gusein-Zade, D. McKenzie, and M. Y. Borodovsky, "Comparison of equations describing the ranked frequency distributions of graphemes and phonemes," *J. Quant. Linguist.*, vol. 3, no. 2, pp. 106–112, 1996.
[36] H. Pande and H. S. Dhami, "Analysis and mathematical modelling of the pattern of occurrence of various *devanāgari* letter symbols according to the phonological inventory of Indic script in Hindi language," *J. Quant. Linguist.*, vol. 22, no. 1, pp. 22–43, 2015.
[37] J. Padgett, "The emergence of contrastive palatalization in Russian," in *Optimality Theory and Language Change*, D. E. Holt, Ed. Kluwer Academic Publishers, 2003, pp. 307–335.
[38] D. C. Hall, "Phonological contrast and its phonetic enhancement: dispersedness without dispersion," *Phonology*, vol. 28, no. 1, pp. 1–54, May 2011.
[39] A. Ashikhmin, A. Barg, and S. Litsyn, "Estimates of the distance distribution of codes and designs," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 1050–1061, Mar. 2001.
[40] J. Liljencrants and B. Lindblom, "Numerical simulation of vowel quality systems: The role of perceptual contrast," *Language*, vol. 48, no. 4, pp. 839–862, Dec. 1972.
[41] H. Cohn and Y. Zhao, "Energy-minimizing error-correcting codes," *IEEE Trans. Inf. Theory*, vol. 60, no. 12, pp. 7442–7450, Dec. 2014.
[42] J. Makhoul, S. Roucos, and H. Gish, "Vector quantization in speech coding," *Proc. IEEE*, vol. 73, no. 11, pp. 1551–1588, Nov. 1985.
[43] J. Yang and J. Y.-T. Leung, "A generalization of the weighted set covering problem," *Nav. Res. Logist.*, vol. 52, no. 2, pp. 142–149, Mar. 2005.
[44] M. A. Changizi and S. Shimojo, "Character complexity and redundancy in writing systems over human history," *Proc.-R. Soc. Lond., Biol. Sci.*, vol. 272, no. 1560, pp. 267–275, Feb. 2005.
[45] Y. Kochman, A. Mazumdar, and Y. Polyanskiy, "Results on combinatorial joint source-channel coding," in *Proc. IEEE Inf. Theory Workshop (ITW'12)*, Sep. 2012, pp. 10–14.