ASR for Under-Resourced Languages from Probabilistic Transcription

Mark Hasegawa-Johnson[1], *Senior Member, IEEE* Preethi Jyothi[1], *Member, IEEE* Daniel McCloy[2], Majid Mirbagheri[2], Giovanni di Liberto[3], Amit Das[1], *Student Member, IEEE* Bradley Ekin[2], *Student Member, IEEE* Chunxi Liu[4], Vimal Manohar[4], *Student Member, IEEE* Hao Tang[5], Edmund C. Lalor[3], Nancy Chen[6], *Senior Member, IEEE* Paul Hager[7], Tyler Kekona[2], Rose Sloan[8], and Adrian KC Lee[2] *Member, IEEE*

1. University of Illinois, 2. University of Washington, 3. Trinity College, Dublin, 4. Johns Hopkins University, 5. Toyota Technological Institute Chicago, 6. Institute for Infocomm Research, 7. MIT, 8. Yale University

Abstract

In many under-resourced languages it is possible to find text, and it is possible to find speech, but transcribed speech suitable for training automatic speech recognition (ASR) is unavailable. In the absence of native transcription, this paper proposes the use of a probabilistic transcription: a probability mass function over possible phonetic transcripts of the waveform. Three sources of probabilistic transcription are demonstrated. First, self-training is a well-established semi-supervised learning technique, in which a cross-lingual ASR first labels unlabeled speech, and is then adapted using the same labels. Second, mismatched crowdsourcing is a recent technique in which non-speakers of the language are asked to write what they hear, and their nonsense transcriptions are decoded using noisy channel models of second-language speech perception. Third, EEG distribution coding is a new technique in which non-speakers of the language listen to it, and their electrocortical response signals are interpreted to indicate probabilities. ASR was trained in four languages with no transcribed training speech. Adaptation using mismatched crowdsourcing significantly outperformed self-training, and both significantly outperformed a cross-lingual baseline. EEG distribution coding and text-derived phone language models were both shown to improve the quality of probabilistic transcriptions derived from mismatched crowdsourcing.

Index Terms

Automatic speech recognition, Under-resourced languages, Mismatched crowdsourcing, EEG

EDICS Category: SPE-RECO

ASR for Under-Resourced Languages from Probabilistic Transcription

I. Introduction

Automatic speech recognition (ASR) has the potential to provide database access, simultaneous translation, and text/voice messaging services to anybody, in any language, dramatically reducing linguistic barriers to economic success. To date, ASR has failed to achieve its potential, because successful ASR requires very large labeled corpora. Current methods require about 1000 hours of transcribed speech per language, transcribed at a cost of about 6000 hours of human labor; the human transcribers must be computer-literate, and they must be native speakers of the language being transcribed. In many languages, the idea of recruiting dozens of computer-literate native speakers is impractical, sometimes even absurd.

Instead of recruiting native transcripts in search of a perfect reference transcription, this paper proposes the use of probabilistic transcriptions. A probabilistic transcription is a probability mass function, $\rho_\Phi(\phi)$, specifying, as a real number between 0 and 1, the probability that any particular phonetic transcription ϕ is the correct transcription of the utterance. Prior to this work, machine learning has almost always assumed that the training dataset contains either deterministic transcriptions ($\rho_{DT}(\phi) \in \{0, 1\}$, commonly called "supervised training") or completely untranscribed utterances (commonly called "unsupervised training," in which case we assume that $\rho_{LM}(\phi)$ is given by some *a priori* language model). This article proposes that, even in the absence of a deterministic transcript, there may be auxiliary sources of information that can be applied to create a probabilistic transcription whose entropy is lower than that of the language model, and that machine learning methods applied to the probabilistic transcription are able to make use of its reduced entropy in order to learn a better speech recognizer. In particular, this paper considers three useful auxiliary sources of information:

1) SELF TRAINING: ASR pre-trained in other languages is used to transcribe unlabeled training data in the target language.
2) MISMATCHED CROWDSOURCING: Human crowd workers who don't speak the target language are asked to transcribe it as if it were a sequence of nonsense syllables.
3) EEG DISTRIBUTION CODING: Humans who do not speak the target language are asked to listen to its extracted syllables, and their EEG responses are interpreted as a probability mass function over possible phonetic transcriptions.

II. BACKGROUND

Consider the problem of developing speech technology in a language with few internet-connected speakers. Suppose we require that, in order to develop speech technology, it is necessary first to have (1) some amount of recorded speech audio, and (2) some amount of text written in the target language. These two requirements can be met by at least several hundred languages: speech audio can be recorded during weekly minority-language broadcasts on a local radio station, and text can be acquired from printed pamphlets and literacy primers. Recorded speech is, however, not usually transcribed; and the requirement of native language transcription is beyond the economic capabilities of many minority-language communities.

A. ASR in Under-Resourced Languages

Krauwer [29] defined an under-resourced language to be one that lacks one or more of: stable orthography, significant presence on the internet, linguistic expertise, monolingual tagged corpora, bilingual electronic dictionaries, transcribed speech, pronunciation dictionaries, or other similar electronic resources. Berment [5] defined a rubric for tabulating the resources available in any given language, and proposed that a language should be called "under-resourced" if it scored lower than 10.0/20.0 on the proposed rubric. By these standards, technology methods for under-resourced languages are most often demonstrated on languages that are not really under-resourced: for example, ASR may be trained without transcribed speech, but the quality of the resulting ASR can only be scientifically proven by measuring its phone error rate (PER) or word error rate (WER) using transcribed speech. The intention, in most cases, is to create methods that can later be ported to truly under-resourced languages.

The International Phonetic Alphabet (IPA [23]) is a set of symbols representing speech sounds (phones) defined by the principle that, if two phones are used in any language to make meaningful linguistic contrasts (i.e., they represent distinct phonemes), then those phones should have distinct symbolic representations in the IPA. This makes the IPA a natural choice for transcriptions used to train cross-language ASR systems, and indeed ASR in a new language can be rapidly deployed using acoustic models trained to represent every distinct symbol in the IPA [43]. However, because IPA symbols are defined phonemically, there is no guarantee of cross-language equivalence in the acoustic properties of the phones they represent. This problem arises even between dialects of the same language: a monolingual Gaussian mixture model (GMM) trained on five hours of Levantine Arabic can be improved by adding ten hours of Standard Arabic data, but only if the log likelihood of cross-dialect data is scaled by 0.02 [20].

Better cross-language transfer of acoustic models can be achieved, but only by using structured transfer learning methods, including neural networks (NN) and subspace Gaussian mixture models (SGMM). SG-

MMs use language-dependent GMMs, each of which is the linear interpolation of language-independent mean and variance vectors [41], e.g., 16% relative WER reduction was achieved in Tamil by combining SGMM with an acoustic data normalization technique [34]. NN transfer learning can be categorized as tandem, bottleneck, pre-training, phone mapping, and multi-softmax methods. In a tandem system, outputs of the NN are Gaussianized, and used as features whose likelihood is computed with a GMM [17]; in a bottleneck system, features are extracted from a hidden layer rather than the output layer. Both tandem [49] and bottleneck [52] features trained on other languages can be combined with GMMs [52] or SGMMs [22] trained on the target language in order to improve WER.

A hybrid ASR is a system in which the NN terminates in a softmax layer, whose outputs are interpreted as phone [38] or senone [9] probabilities. Knowledge of the target language phone inventory is necessary to train a hybrid ASR, but it is possible to reduce WER by first pre-training the NN hidden layers with multilingual data [19], [50]. A hybrid ASR can be constructed using very little in-language speech data by adding a single phone-mapping layer to the output of the multilingual NN; the phone mapping layer can be trained using a small amount of in-language speech data [47], even if context-dependent senones are mapped instead of phones [12]. A multi-softmax system integrates phone mapping into the original training procedure, by training a network with several different language-dependent softmax layers, each of which is the linear transform of a multilingual shared hidden layer. Multi-softmax systems have reduced WER in tandem [42], bottleneck [52], and hybrid [19] ASR.

B. Self-Training

Self-training is a class of semi-supervised learning techniques in which a classifier is first trained on any available labeled examples, then used to classify the unlabeled data. The classifier is then re-trained using its own labels as targets. Self-training is frequently used to adapt ASR from a well-resourced language to an under-resourced language [32], [8], or in some cases, to create target-language ASR by adapting several source-language ASRs [53]. A self-trained classifier tends to be too conservative, because the tails of the data distribution are truncated by the self-labeling process [44]; on the other hand, a self-trained classifier needs to be conservative, because the error rate of the learned classifier increases at a rate more than proportional to the error rate of the self-labeling process [21]. Self-training is therefore most useful when the in-language training data are first filtered, to exclude frames with confidence below a threshold [51], and/or weighted, so that some frames are allowed to influence the learned parameters more than others [18]. Self-training of NN-HMM systems has been shown to be about 50% more effective (1.5 times the error rate reduction) as self-training of GMM-HMM systems [21].

C. Mismatched Crowdsourcing

In [26], a methodology was proposed that bypasses the need for native language transcription: mismatched crowdsourcing sends target language speech to crowd-worker transcribers who have no knowledge of the target language, then uses explicit mathematical models of second language phonetic perception to recover an equivalent phonetic transcription (Fig. 1). Majority voting is re-cast, in this paradigm, as a form of error-correcting code (redundancy coding), which effectively increases the capacity of the noisy channel; interpretation as a noisy channel permits us to explore more effective and efficient forms of error-correcting codes.

Assume that cross-language phone misperception is a finite-memory process, and can therefore be modeled by a finite state transducer (FST). The complete sequence of representations from utterance language to annotation language can therefore be modeled as a noisy channel represented by the composition of up to three consecutive FSTs (Fig. 1): a pronunciation model, a misperception model, and an inverted grapheme-to-phoneme (G2P) transducer. The pronunciation model is an FST representing processes that distort the canonical phoneme string during speech production, including processes of reduction and coarticulation. The misperception model represents the mapping of the uttered phone string (in symbols matching the phone set of the spoken language) to the perceived phone string (in symbols matching the phone set of the annotation language). Finally, the transcriber maps heard phones to nonsense words in the annotation language; the mapping from phones to orthography is an inverted G2P.

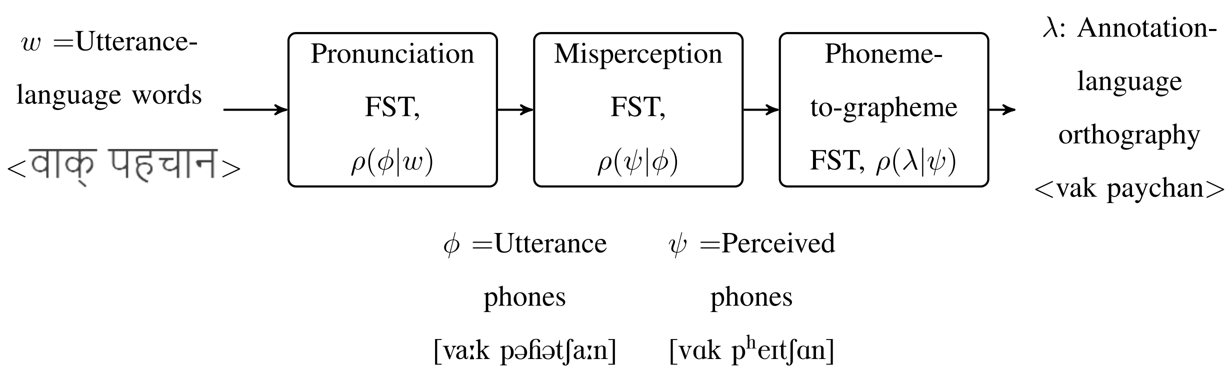

Fig. 1. Mismatched Crowdsourcing: crowd workers on the web are asked to transcribe speech in a language they do not know. Annotation mistakes are modeled by a finite state transducer (FST) model of utterance-language pronunciation variability (reduction and coarticulation), composed with an FST model of non-native speech misperception (mapping utterance-language phones to annotation-language phones), composed with an inverted grapheme-to-phoneme (G2P) transducer.

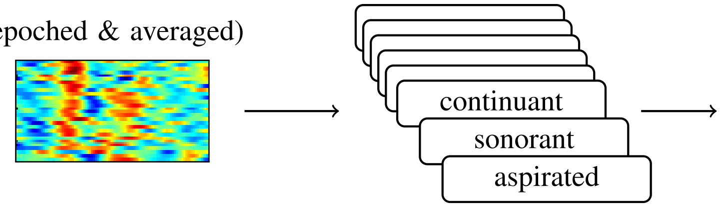

Fig. 2. EEG responses are recorded while listeners hear speech in their native language. For each listener, a bank of distinctive feature classifiers are trained. Listeners then hear speech in an unfamiliar language, and their EEG responses are classified, estimating a listener-language probabilistic transcription of the non-native speech.

D. Electrophysiology of Speech Perception

The human auditory system is sensitive to within-category distinctions in speech sounds, and such pre-categorical perceptual distinctions may be lost in transcription tasks, where listeners must filter their percepts through the limited number of categorical representations available in their native language orthography. EEG distribution coding is a proposed new method that interprets the electrical evoked potentials of untrained listeners (measured by electroencephalography or EEG) as a posterior probability distribution over the phone set of the utterance language (Fig. 2). Transcribers, in this scenario, listen to speech in both their native language and an unfamiliar non-native target language, while their EEG responses are recorded. From their responses to English speech, an English-language EEG phone recognizer is trained [11]. Misperception probabilities $\rho(\psi|\phi)$ are then estimated: for each non-native phone ϕ, the classifier outputs are interpreted as an estimate of $\rho(\psi|\phi)$.

III. Algorithms that Induce a Probabilistic Transcription

A deterministic transcription is a sequence of phone symbols, $\phi^\ell = [\phi^\ell_1, \ldots, \phi^\ell_M]$ where ϕ^ℓ_m is a symbol drawn from the phone set of the utterance language. We assume that ϕ^ℓ_m can be encoded using an IPA symbol [23]. The superscript specifies that ϕ^ℓ is the transcription of the ℓ^{th} waveform in a database; the collection of all transcriptions is $\phi = \left\{\phi^1, \ldots, \phi^L\right\}$.

A probabilistic transcription is a probability mass function (pmf) over the set of deterministic transcriptions. Capital letters denote random variables, lowercase denote instances. Φ^ℓ_m is a random variable whose instance is ϕ^ℓ_m, and whose domain Ω_Φ is the union of the set of IPA symbols with the null symbol ($\emptyset$), thus $\Omega_\Phi = \{\ \emptyset$, [a],[i],[ʌ],[æ],$\ldots\}$, with cardinality $|\Omega_\Phi|$ equal to one plus the number of distinct IPA symbols. Similarly, Φ^ℓ is a random variable whose domain is Ω^*_Φ, the set of all sequences composed of

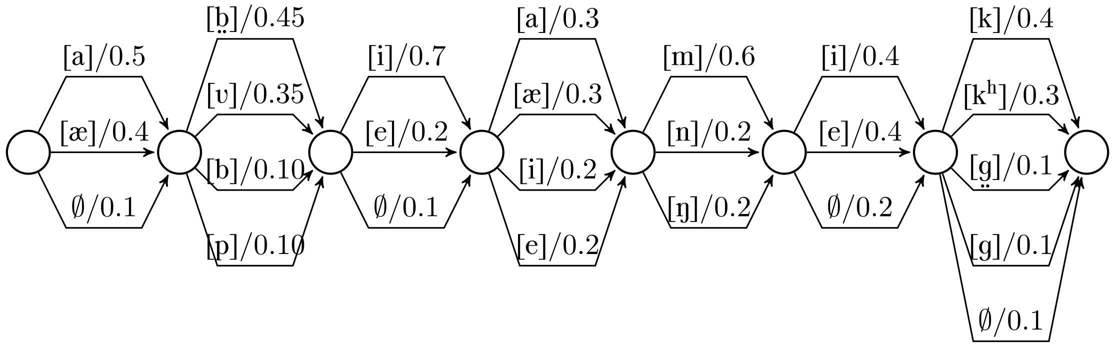

Fig. 3. A probabilistic transcription (PT) is a probability mass function (pmf) over candidate phonetic transcriptions. All PTs considered in this paper can be expressed as confusion networks, thus, as sequential pmfs over the null-augmented space of IPA symbols. In this schematic example, ∅ is the null symbol, symbols in brackets are IPA, and numbers indicate probabilities.

symbols in Ω_Φ. Denote the probability of transcription ϕ^ℓ as $\rho_{\Phi^\ell}(\phi^\ell)$, where ρ (“reference”) means that $\rho_{\Phi^\ell}(\phi^\ell)$ is a reference distribution—a distribution specified by the probabilistic transcription process, and not dependent on ASR model parameters during training. The distribution label Φ^ℓ is omitted when clear from the instance label, e.g., $\rho(\phi^\ell)$, but $\rho_{\Phi^\ell}(u)$. Superscript denotes waveform index, while subscript denotes frame or phone index. Absence of either superscript or subscript denotes a collection, thus $\Phi = \left\{\Phi^1, \ldots, \Phi^L\right\}$ (with instance value $\phi = \left\{\phi^1, \ldots, \phi^L\right\}$) is the random variable over all transcriptions of the database. In all of the work described in this paper, the probabilistic transcription is represented as a confusion network [33], meaning that it is the product of independent symbol pmfs $\rho(\phi^\ell_m)$:

$$\rho(\phi) = \prod_{\ell=1}^{L} \rho(\phi^\ell) = \prod_{\ell=1}^{L} \prod_{m=1}^{M} \rho(\phi^\ell_m) \tag{1}$$

The pmf $\rho(\phi^\ell)$ can be represented as a weighted finite state transducer (wFST) in which edges connect states in a strictly left-to-right fashion without skips, and in which the edges connecting state m to state $m+1$ are weighted according to the pmf $\rho(\phi^\ell_m)$ (Fig. 3).

Three different experimental sources were tested for the creation of a PT. Self-training is now well-established in the field of under-resourced ASR; we adopted the algorithm of Vesely, Hannemann and Burget [51]. Mismatched crowdsourcing used original annotations collected using published methods [27]. EEG was not used independently here, but rather, was used to learn a misperception model applicable to the interpretation of mismatched crowdsourcing.

A. Self Training

The first set of PTs is computed using NN-HMM self-training. The Kaldi toolkit [40] is first used to train a cross-lingual baseline ASR, using training data drawn from six languages not including the target language. The goal of self-training, then, is to adapt the NN-HMM to a database containing L speech waveforms in the target language, each represented by acoustic feature matrix $x^\ell = [x_1^\ell, \ldots, x_T^\ell]$, where x_t^ℓ is an acoustic feature vector. The feature matrix x^ℓ represents an utterance of an unknown phone transcription $\phi^\ell = [\phi_1^\ell, \ldots, \phi_M^\ell]$ which, if known, would determine the sequence but not the durations of senones (HMM states) $s^\ell = [s_1^\ell, \ldots, s_T^\ell]$.

The feature matrix x^ℓ is decoded using the cross-lingual baseline ASR, generating a phone lattice output. Using scripts provided by previous experiments [51], the phone lattice is interpreted as a set of posterior senone probabilities $\rho(s_t^\ell | x_t^\ell)$ for each frame, and the senone posteriors serve as targets for re-estimating the neural network weights. Experiments using other datasets found that self-training should use best-path alignment to specify a binary target for NN training [51], but, apparently because of differences in the adaptation set between our experiments and previous work, we achieve better performance using real-valued targets

B. Mismatched Crowdsourcing

The second set of PTs were computed by sending audio in the target language to non-speakers of the target language, and asking them to write what they hear. It would be preferable to recruit transcribers who speak a language with predictable orthography, but since transcribers in those languages were not readily

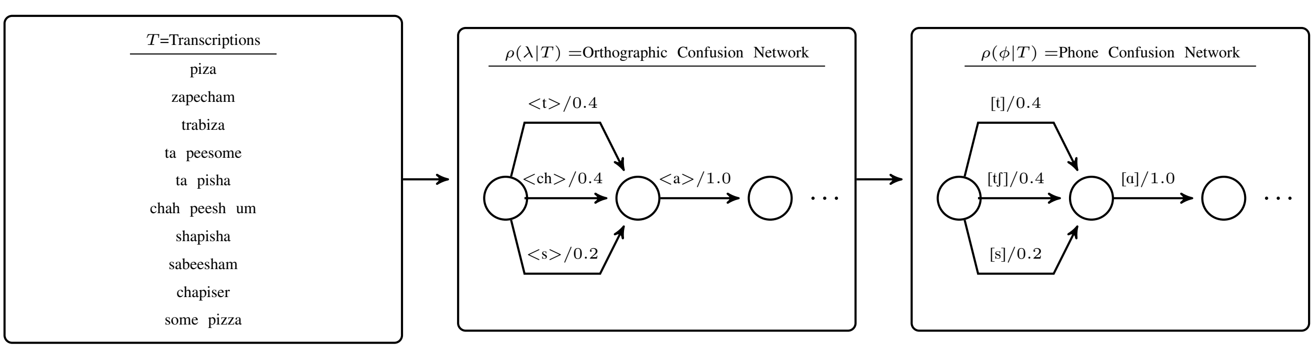

Fig. 4. Probabilistic transcription from mismatched crowdsourcing: Transcripts T are filtered to remove outliers, and merged to create a confusion network over orthographic symbols, $\rho(\lambda|T)$, from which the probabilistic transcription $\rho(\phi|T)$ is inferred. Example shown: Swahili speech, English-speaking transcribers. Symbols in $<>$ are graphemes, symbols in [] are phones, numbers are probabilities.

available, this experiment instead recruited transcribers who speak American English. Denote using T the set of text transcripts produced by these English-speaking crowd workers. Mismatched transcripts must be converted into the form of a pmf over target-language phone sequences, $\rho(\phi|T)$. As an intermediate step towards this goal, prior work [27] developed techniques to merge the transcripts in T into a confusion network $\rho(\lambda|T)$ over representative crowd-worker transcripts, denoted λ (Fig. 4). Formation of $\rho(\lambda|T)$ involves data filtering to remove outliers (based on pair-wise string edit distance among transcripts), expansion of the orthography to an alphabet that includes single-character symbols for digraphs and sequences commonly used to represent single phonemes in English orthography (<ai, ay, ee, oo, ou, aw, ow, bh, ch, dh, gh, jh, kh, ph, sh, th, wh, zh, ck>, and any vowel followed by a word-final silent <e>), and a weighted voting scheme in which the weight of each transcript is proportional to the frequency with which it matches the other transcripts.

Once transcripts have been aligned and filtered to create the orthographic confusion network $\rho(\lambda|T)$, they are then translated into a distribution over phone transcriptions according to:

$$\begin{aligned}\rho(\phi|T) &\approx \max_{\lambda} \rho(\phi|\lambda)\rho(\lambda|T) \\ &= \max_{\lambda} \left(\frac{\rho(\lambda|\phi)}{\rho(\lambda)}\rho(\phi)\right)\rho(\lambda|T) \qquad (2)\end{aligned}$$

The terms other than $\rho(\lambda|T)$ in Equation (2) are estimated as follows. $\rho(\lambda)$ is modeled using a simple context-free prior over the letter sequences in λ. $\rho(\phi)$ is modeled using a bigram phone language model. $\rho(\lambda|\phi)$ is called the misperception G2P, as it maps to graphemes in the annotation language, λ, from phones in the utterance language, ϕ. Section III-C describes methods that decompose $\rho(\lambda|\phi)$ into separate misperception and G2P transducers, but it can also be trained directly using representative transcripts λ (and their corresponding native transcripts) for speech *in languages other than the target language*. We assume that misperceptions depend more heavily on the annotation language than on the utterance language, and that therefore a model $\rho(\lambda|\phi)$ trained using a universal phone set for ϕ is also a good model of $\rho(\lambda|\phi)$ for the target language. Note that, while this assumption is not entirely accurate, it is necessitated by the requirement that no native transcriptions in the target language can be used in building any part of our system.

C. *Estimating Misperceptions from Electrocortical Responses*

The misperception G2P described in Section III-B was estimated using a combination of mismatched and deterministic transcripts of non-target languages. However, with a small amount of transcribed data in the utterance language, it is possible to estimate the misperception G2P using electrocortical measurements of non-native speech perception. In this approach, the misperception G2P is decomposed into two separate transducers, a misperception transducer $\rho(\psi|\phi)$, and an annotation-language G2P $\rho(\lambda|\psi)$:

$$\rho(\lambda|\phi) \approx \sum_{\psi} \rho(\lambda|\psi)\rho(\psi|\phi) \tag{3}$$

where ϕ is a phone string in the utterance language, ψ is a phone string in the annotation language, and λ is an orthographic string in the annotation language. $\rho(\lambda|\psi)$ is an inverted G2P in the annotation language, e.g., trained on the CMU dictionary of American English pronunciations [30]. $\rho(\psi|\phi)$ is the mismatch transducer, specifying the probability that a phone string ϕ in the utterance language will be mis-heard as the annotation-language phone string ψ.

In principle, the mismatch transducer could be computed empirically from a phone confusion matrix, if experimental data on phone confusions were available for all phones in the target language, and those data were based on responses from listeners with the same language background as the crowd worker transcribers. These goals are hard to meet. An alternative is to use distinctive feature representations (originally proposed to characterize the perceptual and phonological natural classes of phonemes [24]) to predict misperceptions based on differences between the distinctive feature values of annotation- and utterance-language phones. Given the assumption that every distinctive feature shared by phones ϕ and ψ independently increases their confusion probability, their confusion probability can be expressed as

$$\rho(\psi|\phi) \propto \exp\left(-\sum_{k=1}^{K} w_k(\phi,\psi)\right) \tag{4}$$

where $w_k(\phi,\psi)$ is zero unless ϕ and ψ share the k^{th} distinctive feature. The assumption of independence is a simplifying assumption, given that many distinctive features have overlapping acoustic correlates. For example, the frequencies of the *two* lowest resonances of the vocal tract (the primary cues for vowel identity) are determined by articulatory gestures of the lips, jaw and tongue that are commonly represented by *three or more* distinctive features (e.g., height, backness, rounding, and advanced tongue root). Moreover, the weights w_k will probably also depend on properties of the speaker and listener (language, dialect, and idiolect), but data to train such a rich model do not exist.

However, a reasonable approximate model can be learned by assuming that w_k depend only on information about the listener, which can be incorporated via measurements of electrocortical activity. In particular, the weights w_k of the distinctive features can be set based on similarity of electrocortical responses (measured using EEG) as determined by a classifier trained on distinctive feature representations and electrocortical responses to the listener's native language phones. Thus, given a set of EEG response signals y recorded when a listener hears audio corresponding to phone ϕ in the annotation language, whose k^{th} distinctive feature is $f_k(\phi)$, and supposing that $g_k(y)$ is the output of a binary classifier trained to detect $f_k(\phi)$ based on speech in the listener's native language[11], then the contributions in Eq. (4) can be estimated as

$$w_k = -\ln \Pr\left\{g_k(y) = f_k(\phi)\right\} \tag{5}$$

IV. Algorithms for Training ASR Using Probabilistic Transcription

An automatic speech recognizer (ASR) is a parameterized probability mass function, $\pi(x, s|\phi, \theta)$, specifying the dependence of acoustic features, x, and senones, s, on the phone transcription ϕ and the parameter vector θ, where the notation $\pi(\cdot)$ denotes a pmf dependent on the ASR parameter vector. Assume a hidden Markov model [3], therefore

$$\pi(x, s|\phi, \theta) = \prod_{\ell=1}^{L}\prod_{t=1}^{T} \pi(s_t^\ell|s_{t-1}^\ell, \phi^\ell, \theta)\pi(x_t^\ell|s_t^\ell, \phi^\ell, \theta)$$

A. Maximum Likelihood Training

Consider two observation-conditional sequence distributions $\pi(s, \phi|x, \theta)$ and $\pi(s, \phi|x, \theta')$, with parameter vectors θ and θ' respectively. The cross-entropy between these distributions is:

$$H\left(\theta\|\theta'\right) = -\sum_{s,\phi} \pi(s, \phi|x, \theta) \ln \pi(s, \phi|x, \theta') \tag{6}$$

$$= \sum_{s,\phi} \pi(s, \phi|x, \theta) \left(\ln \pi(x|\theta') - \ln \pi(x, s, \phi|\theta')\right) \tag{7}$$

$$= \mathcal{L}\left(\theta'\right) - Q\left(\theta, \theta'\right) \tag{8}$$

where the data log likelihood, $\mathcal{L}(\theta')$, and the expectation maximization (EM) quality function, $Q(\theta, \theta')$ [10], are defined by

$$\mathcal{L}\left(\theta'\right) = \ln \pi(x|\theta') \tag{9}$$

$$Q\left(\theta, \theta'\right) = \sum_{s,\phi} \pi(s, \phi|x, \theta) \ln \pi(x, s, \phi|\theta') \tag{10}$$

The Kullback-Leibler divergence between $\pi(s, \phi|x, \theta)$ and $\pi(s, \phi|x, \theta')$ is $D(\theta\|\theta') = H(\theta\|\theta') - H(\theta\|\theta)$. Since $D(\theta\|\theta') \geq 0$ [45],

$$\mathcal{L}\left(\theta'\right) - \mathcal{L}(\theta) \geq Q\left(\theta, \theta'\right) - Q(\theta, \theta) \tag{11}$$

Given any initial parameter vector θ_n, the EM algorithm finds $\theta_{n+1} = \mathrm{argmax}_{\theta'} Q(\theta_n, \theta')$, thereby maximizing the minimum increment in $\mathcal{L}(\theta)$. For GMM-HMMs, the quality function $Q(\theta, \theta')$ is concave and can be analytically maximized; for NN-HMMs it is non-concave, but can be maximized using gradient ascent [4].

The probability $\pi(x, s, \phi|\theta)$ is computed by composing the following three weighted FSTs:

$$H : s^\ell \to s^\ell/\pi(x^\ell|s^\ell, \phi^\ell, \theta) \tag{12}$$

$$C : s^\ell \to \phi^\ell/\pi(s^\ell|\phi^\ell, \theta) \tag{13}$$

$$PT : \phi^\ell \to \phi^\ell/\rho(\phi^\ell) \tag{14}$$

where the notation has the following meaning. The probabilistic transcription, PT, is an FST that maps any phone string $\phi^\ell \in \Omega_\Phi^*$ to itself. This mapping is deterministic and reflexive, but comes with a path cost determined by the transcription probability $\rho(\phi^\ell)$, as exemplified in Fig. 3. The context transducer, C, maps any senone sequence s^ℓ to a phone sequence ϕ^ℓ [35]. This mapping is non-deterministic, and the path cost is determined by the HMM transition weights $a_{ij} = \pi_{S_t^\ell|S_{t-1}^\ell}(j|i, \phi^\ell, \theta)$:

$$\pi(s^\ell|\phi, \theta) = \prod_{\ell=1}^{L}\prod_{t=1}^{T} a_{s_{t-1}^\ell s_t^\ell} \tag{15}$$

The acoustic model, H, maps any senone sequence to itself. This mapping is deterministic and reflexive, but comes with a path cost determined by the acoustic modeling probability

$$\pi(x^\ell|s^\ell, \phi^\ell, \theta) = \prod_{\ell=1}^{L}\prod_{t=1}^{T} \pi(x_t^\ell|s_t^\ell, \theta) \tag{16}$$

The joint probability $\pi(x^\ell, s^\ell, \phi^\ell|\theta)$ is computed by composing the FSTs, then finding the total cost of the path through $H \circ C \circ PT$ with input string s^ℓ and output string ϕ^ℓ. The posterior probability $\pi(s^\ell, \phi^\ell|x^\ell, \theta)$ is computed by pushing the composed FST, then finding the total cost of the path through push $(H \circ C \circ PT)$. Computing the analytical maximum or gradient of $Q(\theta, \theta')$ requires summation over all possible state alignments $s \in \Omega_S^*$. The summation can be performed efficiently using the Baum-Welch algorithm, but experimental tests reported in this paper did not do so, for reasons described in the next subsection.

B. Segmental K-Means Training

The previous subsection demonstrates that $\mathcal{L}(\theta')$ can be increased, at each step of the EM algorithm, by maximizing $Q(\theta, \theta')$. Though $Q(\theta, \theta') - Q(\theta, \theta)$ is a lower bound on $\mathcal{L}(\theta') - \mathcal{L}(\theta)$, Q has properties that make it undesirable as an optimizer for $\mathcal{L}$. Suppose, as often happens, that there is a poor phone sequence, ϕ^p, that is highly unlikely given the correct parameter vector θ^*, meaning that $\pi(x, s, \phi^p|\theta^*)$ is very low. Suppose that the initial parameter vector, θ, is less discriminative, so that $\pi(x, s, \phi^p|\theta) > \pi(x, s, \phi^p|\theta^*)$. In this case $Q(\theta, \theta^*)$ is dominated by the term $\pi(x, s, \phi^p|\theta) \ln \pi(x, s, \phi^p|\theta^*)$, therefore θ^* will never show up as the optimizer of $Q(\theta, \theta')$. Indeed, the best speech recognizer is a parameter vector θ^* that completely rules out poor transcriptions, setting $\pi(x, s, \phi^p|\theta^*) = 0$; but in this case $Q(\theta, \theta^*) = -\infty$, so the EM algorithm can never find a parameter vector θ^* that sets to zero the probability of a poor transcription.

Deterministic transcription does not have this problem, because the transcription specifies the phone sequence. With probabilistic transcription, however, the problem is quite common: if the human transcribers fail to rule out ϕ^p (e.g., because the correct and incorrect transcriptions are perceptually indistinguishable in the language of the transcribers), then the EM algorithm will also never learn to rule out ϕ^p. EM is unable to learn zero-valued probabilities.

EM's inability to learn zero-valued probabilities can be ameliorated by using the segmental K-means algorithm [25], which bounds $\mathcal{L}(\theta')$ as $\mathcal{L}(\theta') - \mathcal{L}(\theta) \geq R(\theta, \theta') - \mathcal{L}(\theta)$, where

$$R(\theta, \theta') = \ln \pi(x, s^*(\theta), \phi^*(\theta)|\theta') \tag{17}$$

$$s^*(\theta), \phi^*(\theta) = \operatorname*{argmax}_{s,\phi} \pi(s, \phi|x, \theta) \tag{18}$$

Given an initial parameter vector θ, therefore, it is possible to find a new parameter vector θ' with higher likelihood by computing its maximum-likelihood senone sequence and phone sequence $s^*(\theta), \phi^*(\theta)$, and by maximizing θ' with respect to $s^*(\theta)$ and $\phi^*(\theta)$. In practice, maximizing $R(\theta, \theta')$ rather than $Q(\theta, \theta')$ is useful for probabilistic transcription because it reduces the importance of poor phonetic transcriptions.

C. *Using a Language Model During Training*

A PT contains significant information beyond any single transcript extracted from the PT. Motivated by this, the statistics for ASR adaptation are accumulated from a lattice derived from the cascade $H \circ C \circ PT$, rather than reducing the PT to its single best path. Though it is disadvantageous to reduce a PT to its best path, it is nevertheless advantageous to incorporate as much information as possible about the target language during adaptation. Define G to be an FST representing the modeled phone bigram probability $\pi(\phi^\ell|\theta) = \prod_{m=1}^{M} \pi(\phi^\ell_m|\phi^\ell_{m-1}, \theta)$. Training results can be improved by using $H \circ C \circ PT \circ G$ to compute segmental K-means.

By assumption, phone bigram information is not available from speech: we assume that there is no transcribed speech in the target language. A reasonable proxy, however, can be constructed from text. Fig. 5 shows text data downloaded from Wikipedia in Swahili, and a segment of a rule-based, character-by-character G2P for the Swahili language [16]. By passing the former through the latter, it is possible to generate synthetic phoneme sequences in the target language.

Composing $PT \circ G$ is complicated by the presence of null transitions in the PT. A null transition in the PT matches a non-event in the language model, for which normal FST notation has no representation. In

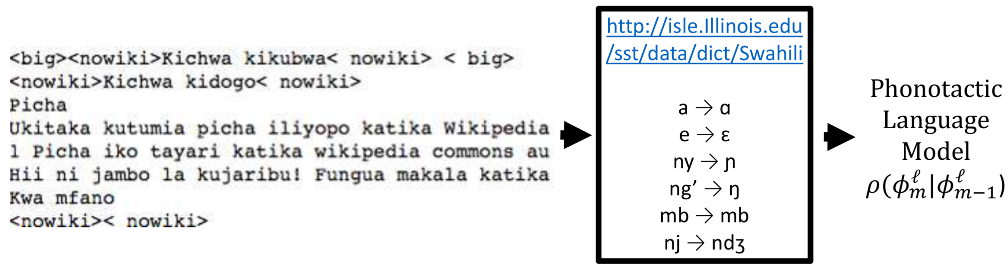

Fig. 5. A phonotactic language model (a bigram language model over phone sequences) can be trained using text data downloaded from Wikipedia (left), then converted into phone strings in the target language using a simple character-based grapheme-to-phoneme transducer (center). In this example, the target language is Swahili.

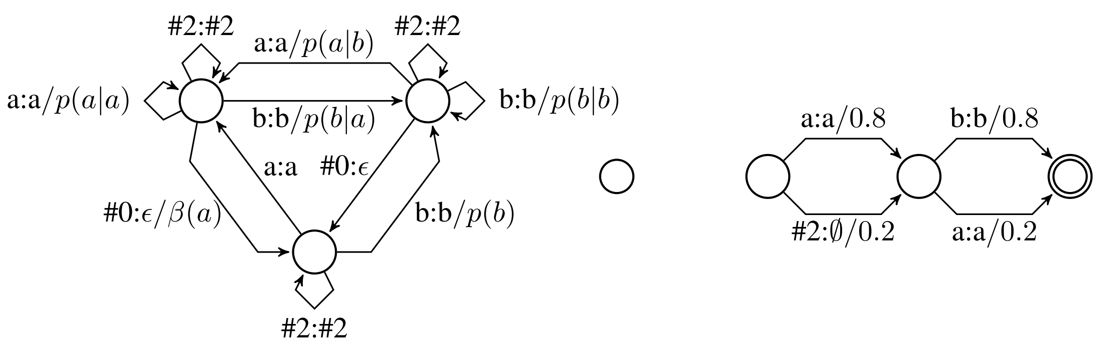

Fig. 6. Deletion edges in the probabilistic transcript (edges with the special null-phone symbol, $\emptyset$), required special handling in order to use information from a phonotactic language model. As shown, a new type of null symbol, "#2", was invented to represent the input for every PT edge with an $\emptyset$ output (right). Such edges were only allowed to match with state self-loops, newly added to the language model (left) in order to consume such non-events in the transcript. a,b: regular phone symbols, ϵ: null-string, $p(b|a)$: bigram probability, $\beta(a)$: language model backoff.

order to compose the PT with the language model, therefore, it is necessary to introduce a special type of "non-event" symbol, here denoted "#2", into the language model (Fig. 6). As shown in Fig. 6, a language model "non-event" is a transition that leaves any state, and returns to the same state (a self-loop). Such self-loops, labeled with the special symbol "#2" on both input and output language, are added to every state in the phonotactic language model (left-hand side of Fig. 6). The probabilistic transcript, then, is augmented with the special symbol "#2" as the input-language symbol for every null-output edge (output symbol is $\phi_m^\ell = \emptyset$).

D. *Maximum* A Posteriori *Adaptation*

Training from PTs can be improved by starting from a multilingual ASR, and adapting its parameters to PTs in the target language. The Bayesian framework for maximum *a posteriori* (MAP) estimation has been widely applied to GMM and HMM parameter estimation problems such as parameter smoothing and speaker adaptation [14].

Formally, for an unseen target language, denote its acoustic observations $x = (x_1^1, \ldots, x_T^L)$, and its acoustic model parameter set as θ, then the MAP parameters are defined as:

$$\theta_{\text{MAP}} = \underset{\theta}{\operatorname{argmax}}\, \pi(\theta|x) = \underset{\theta}{\operatorname{argmax}}\, \pi(x|\theta)\pi(\theta) \tag{19}$$

where $\pi(\theta)$ is the product of conjugate prior distributions, centered at the parameters of a cross-lingual baseline GMM-HMM. In a GMM-HMM, the acoustic model is computed by choosing a Gaussian component, G_t^ℓ, whose mixture weight is $c_{jk} = \pi_{G_t^\ell|S_t^\ell}(k|j)$, and whose mean vector and covariance matrix are μ_{jk} and Σ_{jk}. Maximum likelihood trains these parameters by computing $\gamma_t^\ell(j,k) = \pi_{S_t^\ell, G_t^\ell}(j,k|x^\ell,\theta)$, then accumulating weighted average acoustic frames with weights given by $\gamma_t^\ell(j,k)$. Segmental K-means quantizes $\pi_{S_t^\ell}(j|x^\ell,\theta) \rightarrow \{0,1\}$ using forced alignment, then proceeds identically. MAP adaptation assigns, to each parameter, a conjugate prior $\pi(\theta)$ with mode equal to $\bar{\theta}$ (the parameters of the multilingual baseline), and with a confidence hyperparameter τ_θ, resulting in re-estimation formulae that are linearly interpolated between the baseline parameters $\bar{\theta}$ and the statistics of the adaptation data, for example:

$$c'_{jk} = \frac{\tau_c \bar{c}_{jk} + \sum_{\ell,t} \gamma_t^\ell(j,k)}{\sum_\kappa \left(\tau_c \bar{c}_{j\kappa} + \sum_{\ell,t} \gamma_t^\ell(j,\kappa)\right)} \tag{20}$$

E. Neural Networks

The NN acoustic model is $\pi_{X_t^\ell|S_t^\ell}(v|j,\theta) \propto y_t^\ell(j)$,

$$y_t^\ell(j) = \frac{1}{c_j} \frac{\exp\left(w_j^T h_t(v, w_{vh})\right)}{\sum_k \exp\left(w_k^T h_t(v, w_{vh})\right)} \tag{21}$$

whose parameters $\theta = \{c_j, w_j, w_{vh}\}$ include the senone priors c_j, the softmax weight vectors w_j, and the parameters defining the hidden nodes $h_t(v, w_{vh})$. NNs are trained by using a GMM-HMM to compute an initial senone posterior, $\pi_{S_t^\ell}(j|x^\ell,\theta)$, then minimizing the cross-entropy between the estimated senone posterior and the neural network output $y_t^\ell(j)$, using gradient descent in the direction

$$-\nabla_\theta H(S^\ell \| Y^\ell) = \sum_{t=1}^{T} \sum_j \frac{\pi_{S_t^\ell}(j|x^\ell,\theta)}{y_t^\ell(j)} \nabla_\theta y_t^\ell(j) \tag{22}$$

NN training with deterministic transcriptions is improved by quantizing $\pi_{S_t^\ell}(j|x^\ell,\theta) \rightarrow \{0,1\}$ using forced alignment [38]. Preliminary experiments showed that forced alignment also improves the accuracy of NNs trained from probabilistic transcriptions: the best path through the PT, and the best alignment of the resulting senones to the waveform, were both computed using forced alignment. The resulting best senone string was used to train a NN using Eq. (22).

V. Experimental Methods

ASR was trained in four target languages, each using no native transcriptions. Unspoken texts and untranscribed audio were acquired in each language (Sec. V-A). Texts from mismatched crowdsourcing were decoded using a multilingual misperception G2P (Sec. V-B), or using a language-specific misperception G2P estimated using EEG (Sec. V-C). Baseline cross-lingual systems were trained using native transcriptions from several different languages not including the target language (Section V-D). Finally, parameters of the acoustic model were adapted using PTs in the target language, derived from either self-training or from mismatched transcriptions (Sec. V-E).

A. Data

Speech data were extracted from publicly available podcasts [48] hosted in 68 different languages. In order to generate test corpora (in which it is possible to measure phone error rate), advertisements were posted at the University of Illinois seeking native speakers willing to transcribe speech in any of these 68 languages. Of the ten transcribers who responded, six people were each able to complete one hour of speech transcription (the other four dropped out). One additional language was transcribed by workers recruited at I^2R in Singapore, yielding a total of seven languages with native transcriptions suitable for testing an ASR: Arabic (arb), Cantonese (yue), Dutch (nld), Hungarian (hun), Mandarin (cmn), Swahili (swh) and Urdu (urd).

The podcasts were not entirely homogeneous in the target language and contain utterances interspersed with segments of music and English. A simple GMM-based language identification system was developed as a first pass over the podcasts in order to isolate regions that correspond mostly to the target language. These long segments were then split into smaller $\approx$ 5-second segments. This was to enable easy labeling by the native transcribers, and more importantly to allow for the collection of mismatched transcriptions that required the speech segments to be short. To further check that only speech clips in the target language were retained, the native transcribers were asked to omit any 5-second clips that contained music, significant amounts of noise, English speech or speech from multiple speakers. The resulting transcribed speech clips roughly amounted to 45 minutes of speech in Urdu and 1 hour of speech in the remaining six languages. The orthographic transcriptions for these clips were then converted into phonemic transcriptions using language-specific dictionaries and G2P mappings (these resources are detailed in Section V-D). For each language, we chose a random 40/10/10 minutes split into training, development and evaluation sets. Table I describes the resulting training, development and evaluation sets.

Language (ISO 639-3)	Speech (# phones) Train	Dev	Eval
arb	32486	8208	8191
yue	32693	6860	8638
nld	27314	6943	6582
hun	29461	7873	7474
cmn	29461	8244	7035
swh	28571	7658	7441
urd	21275	5808	3689

TABLE I

DATA STATISTICS, SEVEN LANGUAGES, # PHONES IN THE TRAIN/DEV/EVAL SETS.

B. *Mismatched Crowdsourcing*

Mismatched transcriptions were collected from crowd workers (Turkers) on Amazon Mechanical Turk [1] for all the data listed in Table I. The crowdsourcing task setup is described in [27]. Each 5-sec speech segment was further split into 4 non-overlapping segments to make the non-native listening task easier. The crowdsourcing task was set up as described in [27]; briefly, the short segments were played to Turkers, who transcribed what they heard (typically in the form of nonsense syllables) using English orthography. Each short recording segment was transcribed by 10 distinct Turkers. More than 2500 Turkers participated in these tasks, with roughly 30% of them claiming to know only English. (Spanish, French, German, Japanese, Chinese were some of the other languages listed by the Turkers.)

C. *EEG Recording and Analysis*

To compute distinctive feature weights used in estimating the misperception transducer as shown in Eqs. (4) and (5), recordings of cortical activity in response to non-native phones were made using EEG. Signals were acquired using a BrainVision actiCHamp system with 64 channels and 1000 Hz sampling frequency. All procedures were approved by the University of Washington Institutional Review Board.

Auditory stimuli were consonant-vowel (CV) syllables representing consonants of three languages: English, Dutch and Hindi. The inclusion of only two non-English languages was dictated by the relatively high number of repetitions needed for good signal-to-noise ratio from averaged EEG recordings. The choice of Dutch and Hindi was made based on language phonological similarity, defined as the number of many-to-one mappings ($N_{M2O}(\Omega_\Phi)$) between the English phoneme inventory (Ω_Ψ) and the non-English phoneme inventory Ω_Φ. Using distinctive feature representations of the phonemes in each inventory from

Ω_Φ	$N_{M2O}(\Omega_\Phi)$	Ω_Φ	$N_{M2O}(\Omega_\Phi)$	Ω_Φ	$N_{M2O}(\Omega_\Phi)$
spa	0.862	yue	1.280	cmn	1.531
por	1.152	jpn	1.333	amh	1.844
nld	1.182	vie	1.393	hun	1.857
deu	1.258	kor	1.429	hin	2.848

TABLE II

FREQUENCY OF MANY-TO-ONE MAPPINGS $N_{M2O}(\Omega_\Phi)$ BETWEEN PHONEME INVENTORY Ω_Φ AND THE INVENTORY OF ENGLISH. LANGUAGES ARE REPRESENTED BY THEIR ISO 639-3 CODES.

the PHOIBLE database [37], a many-to-one mapping was defined by finding, for each non-English phoneme ϕ, the English phoneme $\psi^*(\phi)$ to which it is most similar:

$$\psi^*(\phi) = \operatorname{argmin} \sum_k \delta\left(f_k(\psi) \neq f_k(\phi)\right) \tag{23}$$

The number of many-to-one collisions is then defined as

$$N_{M2O}(\Omega_\Phi) = \frac{1}{|\Omega_\Psi|} \sum_{\phi_1 \neq \phi_2} \delta\left(\psi^*(\phi_1) = \psi^*(\phi_2)\right) \tag{24}$$

where $|\Omega_\Psi|$ is the size of the English phoneme inventory. The frequency of many-to-one mappings is listed in Table II for several languages.

Based on this criterion, Hindi was chosen for having a large number of many-to-one mappings with English, while Dutch has relatively few. Many-to-one mappings are expected to pose a problem for the non-native transcription task being modeled by the misperception transducer, so to test the contribution of EEG we chose languages that differed greatly in this property. Note that, although Hindi podcasts were not included in the training data described in Section V-A, colloquial spoken Hindi and Urdu are extremely similar phonologically [28], and considering that the auditory stimuli for the EEG portion of this experiment are simple CV syllables, it is reasonable to consider Hindi and Urdu as equivalent for the purpose of computing feature weights for the misperception transducer.

To construct the auditory stimuli, two vowels and several consonants were selected from the phoneme inventory of each language (18 consonants for English, 17 for Dutch, and 19 for Hindi). Consonants were chosen to emphasize differences in the many-to-one relationships between English-Dutch and English-Hindi, while maintaining roughly equal numbers of consonants for each language. The consonants chosen for each language are given in Table III; the vowels chosen were the same for all three languages (/a/ and /e/).

Two native speakers of each language (one male and one female) were recorded (44100 Hz sampling frequency, 16 bit depth) speaking multiple repetitions of the set of CV syllables for their language.

	Consonants used in the EEG experiment									
eng	p	t	k	tʃ	v	ð	z	m	n	
nld	p	t	ɣ		v		z	m	n	
hin	p b	t̪ d̪ ʈ ɖ	k g		ʋ			m	n̪	ɳ
eng	pʰ	tʰ	kʰ	tʃʰ	f	θ	ʃ	l	ɹ	
nld	pʰ	tʰ	kʰ	tʃʰ	f		ʃ	l	ʀ	j
hin	bʱ	t̪ʰ ʈʰ d̪ʱ ɖʱ	kʰ gʱ							

TABLE III

CONSONANT PHONES USED IN THE EEG EXPERIMENT REPRESENTED USING IPA. VERTICAL ALIGNMENT OF CELLS SUGGESTS MANY-TO-ONE MAPPINGS EXPECTED BASED ON DISTINCTIVE FEATURE VALUES.

Three tokens of each unique syllable were excised from the raw recordings, downsampled to 24414 Hz, and RMS normalized. Recorded syllables had an average duration of 400 ms, and were presented via headphones to one monolingual American English listener. The stimuli were presented in 9 blocks of 15 minutes per block, for a total of 135 minutes. Syllables were presented in random order with an inter-stimulus interval of 350 ms. 21 repetitions of each syllable were presented, for a grand total of 9072 syllable presentations.

EEG recordings were divided into 500 ms epochs. The epoched data were coded with a subset of distinctive features that minimally defined the phoneme contrasts of the English consonants. Where more than one choice of features was sufficient to define those contrasts, preference was given to features that reflect differences in temporal as opposed to spectral features of the consonants, due to the high fidelity of EEG at reflecting temporal envelope properties of speech [11]. The final set of features chosen was: continuant, sonorant, delayed release, voicing, aspiration, labial, coronal, and dorsal.

D. Cross-Lingual Baselines

The goal of building a cross-lingual system is two-fold. One is to define a baseline for generalizing to an unseen language without any labeled audio corpus. The other is have the baseline serve as a starting point for adaptation.

The dataset consists of 40 minutes of labeled audio for training, 10 minutes for development, 10 minutes for testing for each language. The orthographic transcriptions are converted into phonetic transcriptions in the following steps. Beginning with a list of the IPA symbols used in canonical descriptions of all seven languages, any symbol appearing in only one language was merged with a different symbol differing by only one distinctive feature; this process proceeded until each remaining phone symbol is represented in at least two languages. English words are identified and converted to phones with an English G2P trained using CMUdict [30]. We take the canonical pronunciation of a word if the word appears in a lexicon, otherwise estimate the word's pronunciation using a G2P. The Arabic dictionary is from the Qatari Arabic Corpus [13], the Dutch dictionary is from CELEX v2 [2], the Hungarian dictionary was provided by BUT [15], the Cantonese dictionary is from I^2R, the Mandarin dictionary is from CALLHOME [7], and the Urdu and Swahili G2Ps were compiled from rule-based descriptions of the orthographic systems in those two languages [16].

Each HMM was trained with data from six languages, tuned (hyperparameters) on the development set of the seventh language, and tested on the evaluation set of the seventh language. The lexicon of the target language was not used during testing, but two types of language-dependent specialization were allowed. In the first type of specialization, the universal phone set was restricted at test time to output only phones in the target language. In the second type of specialization, a target-language phone bigram language model was trained using phone sequences converted from text. The texts were collected from Wikipedia articles linked from the main page of each language crawled once per day over four months. As an oracle experiment, we also train language dependent HMMs for each individual language with 40 minutes of labeled audio.

E. MAP Adaptation to Probabilistic Transcripts

The baseline and the adapted models were implemented using Kaldi [40]. In order to efficiently carry out the required operations on the cascade $H \circ C \circ PT \circ G$, PT was defined as $\mathrm{proj}_{\mathrm{input}}(\widehat{PT})$, where $\widehat{PT}$ is a wFST mapping phone sequences to English letter sequences (Eq. 2), and $\mathrm{proj}_{\mathrm{input}}$ refers to projecting onto the input labels. For the purposes of computational efficiency, the cascade for $\widehat{PT}$ includes an additional wFST restricting the number of consecutive deletions of phones and insertions of letters (to a maximum of 3). Two additional disambiguation symbols [36] were used to determinize these insertions and deletions in $\widehat{PT}$. MAP adaptation for the acoustic model was carried out for a number of iterations (12 for yue & cmn, 14 for hun & swh, with a re-alignment stage in iteration 10).

VI. Experimental Results

This section reports two types of results. First, subsections VI-A and VI-B report improvements in the quality of probabilistic transcription using information acquired from text-based phone language models and EEG signals, respectively. Second, subsections VI-C and VI-D reports the accuracy of cross-lingual ASR and PT-adapted ASR, respectively.

Method	nld	cmn	urd	arb	hun
Universal set	87.4	88.86	97.95	79.04	92.87
Target set	78.12	87.4	87.81	66.39	84.78
Phone bigram	68.0	70.88	64.67	65.29	63.98

TABLE IV

LABEL PHONE ERROR RATE (LPER) OF PROBABILISTIC TRANSCRIPTIONS FOR NLD=DUTCH, CMN=MANDARIN, URD=URDU, ARB=ARABIC, HUN=HUNGARIAN. METHODS: UNIVERSAL PHONE SET, TARGET-LANGUAGE PHONE SET, TEXT-BASED PHONE BIGRAM.

A. *Mismatched Crowdsourcing*

The quality of a probabilistic transcription derived from mismatched crowdsourcing is significantly improved by using a phone language model during the decoding process ($\rho(\phi)$ in Eq. (2)). A crude measure of the quality of the PTs is given by the label phone error rate (LPER), which measures the difference between $\phi^* = \mathrm{argmax}_\phi\, \rho(\phi|T)$ and a native transcription. Phone language models for each target language were computed from Wikipedia texts using the methods described in Sec. IV-C. LPER of the 1-best path through the resulting PTs are shown in Table. IV. As shown, the use of a phonotactic language model, derived from Wikipedia text, reduces PER by about 10% absolute, in each language.

LPER of the 1-best path does not accurately reflect the extent of information in the PTs that can be leveraged during ASR adaptation. Consider, for example, the four Urdu phones [p,$\mathrm{p^h}$,b,b̤]. An attentive English-speaking transcriber must choose between the two letters <p,b> in order to represent any of these four phones. The misperception G2P therefore maps the letters <p,b> into a distribution over the phones [p,$\mathrm{p^h}$,b,b̤]. There is no reason to expect that the maximizer of $\rho(\phi|\lambda)$ is correct, but there is good reason to expect the correct answer to be a member of a short N-best list ($N \leq 4$ phones/grapheme). A fuller picture is therefore obtained by considering a collection of sequences ϕ that are almost as probable as ϕ^* according to our model. Figure 7 shows the trend of phone error rates (for three languages) obtained by using collections ϕ of increasing size, plotted against an entropy estimate of ϕ, e.g., 1 bit of entropy allows two equally probable choices for each phone in ϕ. We note that the phone error rates significantly drop across all languages, staying within 1 bit of entropy per phone, illustrating the extent of information captured by the PTs.

B. *Misperception Transducer Trained Using EEG*

Epoched and feature-coded EEG data *for the English syllables only* were used to train a support vector machine classifier for each distinctive feature. The classifiers were then used (without re-training) to

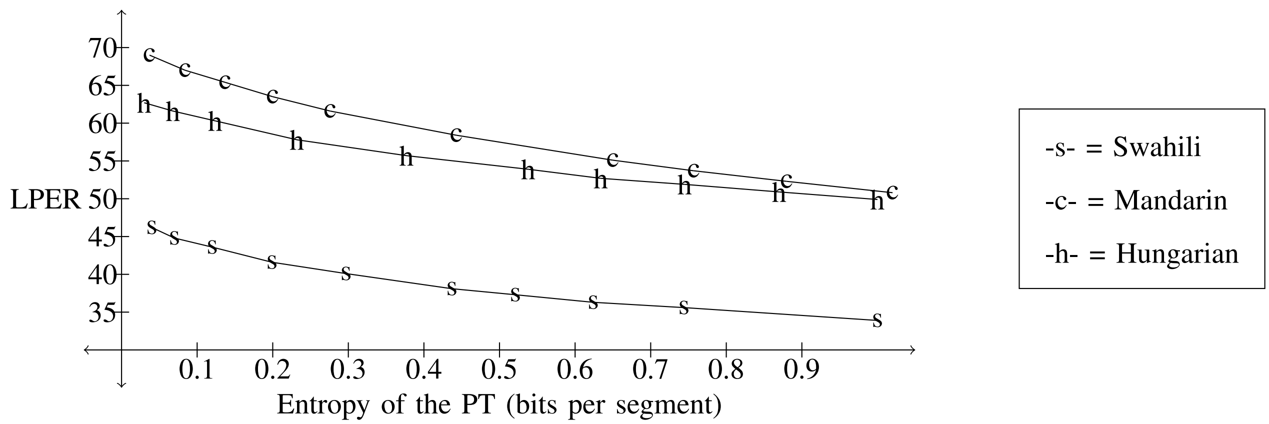

Fig. 7. LPER plotted against entropy rate estimates of phone sequences in three different languages.

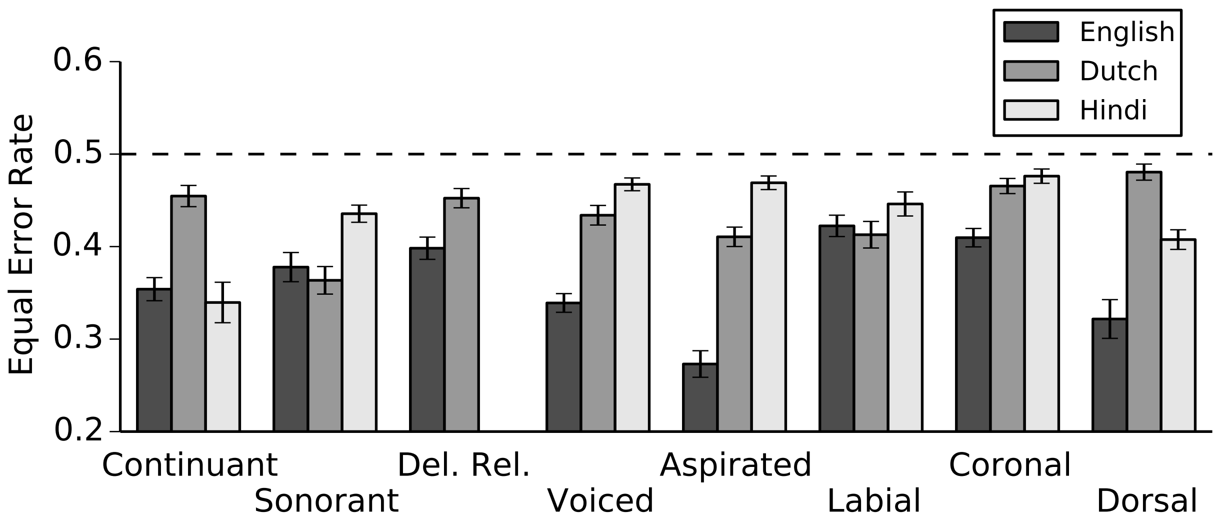

Fig. 8. Classifiers were trained to observe EEG signals, and to classify the distinctive features of the phone being heard. Equal error rates are shown for English (the language used in training; train and test data did not overlap), Dutch, and Hindi. Dashed line shows chance=50%.

classify the EEG responses to the Dutch and Hindi syllables. Fig. 8 shows equal error rates of these classifiers when applied to the three languages.

Eq. (4) defines a log-linear model of $\rho(\psi|\phi)$, the probability that a non-English phoneme ϕ will be perceived as English phoneme ψ. Denote by $\rho_U(\psi|\phi)$ the model of Eq. (4) with uniform weights for all distinctive features. Denote by $\rho_{EEG}(\psi|\phi)$ the same model, but with weights w_k derived from EEG measurements (Eq. (5)). Fig. 9 shows these two confusion matrices: $\rho_U(\psi|\phi)$ on the left, $\rho_{EEG}(\psi|\phi)$ on the right. The entropy of the uniform weighting, $\rho_U(\psi|\phi)$, is too low: when a Dutch phoneme ϕ has a nearest-neighbor $\psi^*(\phi)$ in English, then few other phonemes are considered to be possible confusions.

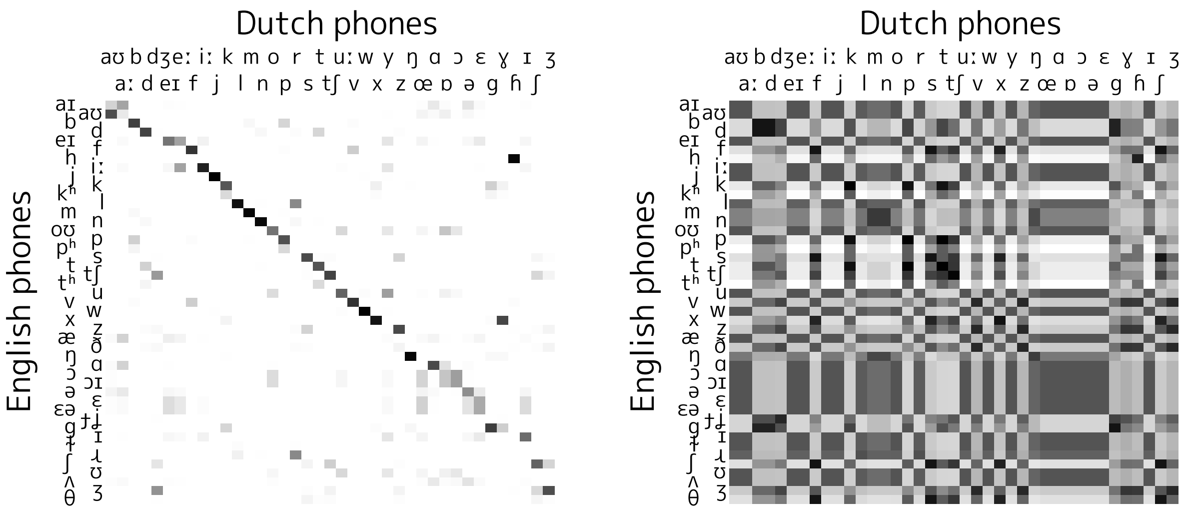

Fig. 9. Phone confusion probabilities between English and Dutch phones using models in which the log probability is proportional to distance between the corresponding distinctive features. Left: all features have the same weight. Right: feature weights equal negative log error rate of EEG signal classifiers.

$\rho_{EEG}(\psi|\phi)$ has a very different problem: since distinctive feature classfiers have been trained for only a small set of distinctive features, there are large groups of phonemes whose confusion probabilities can not be distinguished (giving the figure its block-matrix structure). The faults of both models can be ameliorated by averaging them in some way, e.g., by computing the linear interpolation $\rho_I(\psi|\phi) = \alpha\rho_U(\psi|\phi) + (1-\alpha)\rho_{EEG}(\psi|\phi)$ for some constant $0 \leq \alpha \leq 1$.

In order to evaluate the effectiveness of the EEG-induced misperception transducer we looked at the label phone error rate of mismatched crowdsourcing for the Dutch language when performed using 1) a multilingual misperception model $\rho(\lambda|\phi)$, 2) feature-based misperception transducer computed using uniform weighting, $\rho_U(\psi|\phi)$, or 3) EEG-induced transducer combined with the feature-based transducer, $\rho_I(\psi|\phi)$. To combine the two transducers, the value of the parameter α was optimized on a separate development data set. As shown in Table V, phone error rates were improved 1.5% and 2.5% relative to mismatched crowdsourcing by using the feature-based and combined misperception transducers (respectively).

C. *Cross-Lingual Baseline*

Table VI compares cross-lingual ASR using universal phone set and phone language model, cross-lingual ASR using language-dependent phone set and phone language model, and monolingual ASR.

	multilingual	feature-based	EEG-induced +feature-based
LPER	70.43	69.44	68.61

TABLE V

COMPARISON OF LPERS FOR PROBABILISTIC TRANSCRIPTION OF THE DUTCH EVALUATION SET USING DIFFERENT SCHEMES TO COMPUTE MISPERCEPTION G2PS.

data	acoustic model	language model	yue	hun	cmn	swh
cross-lingual	GMM	cross-lingual	79.64 (79.83)	77.13 (77.85)	83.28 (82.12)	82.99 (81.86)
cross-lingual	NN	cross-lingual	78.62 (77.58)	75.98 (76.44)	81.86 (80.47)	82.30 (81.18)
cross-lingual	GMM	text	68.40 (68.35)	68.62 (66.90)	71.30 (68.66)	63.04 (64.73)
cross-lingual	NN	text	66.54 (65.28)	66.08 (66.58)	65.77 (64.80)	64.75 (65.04)
monolingual	GMM	transcript	32.77 (34.61)	39.58 (39.77)	32.21 (26.92)	35.33 (46.51)
monolingual	NN	transcript	27.67 (28.88)	35.87 (36.58)	27.80 (23.96)	34.98 (41.47)

TABLE VI

PERS OF UNADAPTED CROSS-LINGUAL AND MONOLINGUAL ASR ON THE EVALUATION SETS (DEVELOPMENT SETS ARE IN PARENTHESES). TEXT-BASED LANGUAGE MODELS ARE TRAINED USING WIKIPEDIA. TRANSCRIPT-BASED LANGUAGE MODELS ARE BASED ON NATIVE TRANSCRIPTS OF THE TRAINING DATA.

From the comparison of different baseline systems, we can reach the following conclusions. First, even with only 40 minutes of training data, a NN is able to outperform a GMM. Second, however, the standard speech pipeline performs poorly on unseen languages. Using a language-specific phonotactic language model gives significant improvement over the language-independent phonotactic model, but nevertheless significantly underperforms a system that has seen the test language during training. This is true even if the system has seen closely related languages during training: the Cantonese cross-lingual system has seen Mandarin during training, and the Mandarin system has seen Cantonese during training, but neither system is able to generalize well from its training language to its test language.

D. ASR Trained Using Probabilistic Transcriptions

This section demonstrates that PT adaptation improves the generalization capability of cross-lingual ASR to an unseen target language. Adaptation to ASR-derived PTs (self-training) significantly reduces PER, as has been previously reported [51]. PTs derived from human mismatched crowdsourcing provide significant further PER reduction.

Language Code	Cross-Lingual (CL)	Self-training (ST)	CL + PT adaptation (PT-ADAPT)	% Rel. redn over CL	% Rel. redn over ST
GMM-HMM					
yue	68.40 (68.35)		**57.20 (56.57)**	16.4** (17.1)	10.3** (9.2)
hun	68.62 (66.90)		**56.98 (57.26)**	16.9** (14.3)	10.2** (9.9)
cmn	71.30 (68.66)		**58.21 (57.85)**	18.4** (15.7)	10.3** (9.7)
swh	63.04 (64.73)		**44.31 (48.88)**	29.6** (24.6)	24.7** (18.4)
NN-HMM					
yue	66.59 (65.41)	63.79 (62.46)	**53.64 (53.80)**	19.4** (17.7)	15.9** (13.9)
hun	66.43 (67.18)	63.53 (63.50)	**56.70 (58.45)**	14.6** (13.0)	10.8** (8.0)
cmn	65.77 (64.80)	64.90* (64.00)	**54.07 (53.13)**	17.8** (18.0)	16.7** (17.0)
swh	65.30 (65.11)	58.76** (59.81)	**44.73 (48.60)**	31.5** (25.4)	23.9** (18.7)

TABLE VII

PERS ON THE EVALUATION AND DEVELOPMENT SETS (DEVELOPMENT IN PARENTHESES) BEFORE AND AFTER ADAPTATION WITH PTS. MAPSSWE SIGNIFICANCE TESTING: * MEANS $p \leq 0.003$, ** MEANS $p < 0.001$.

Table VII presents phone error rates (PERs) on the evaluation (and development) sets for four different languages. The column titled CL lists cross-lingual baseline error rates, copied from Table VI.

In Table VII, the column labeled ST lists the PERs of self-trained ASR systems. Self-training was only performed using NN systems; no self-training of GMMs was performed, because previous studies [19] reported it to be less effective. Differences between the evaluation set PERs of ST and CL systems were tested for statistical significance using the MAPSSWE test of the `sc_stats` tool [39]. There are 20 independent statistical comparisons in Table VII; the study-corrected significance level of $0.05/20 = 0.0025$ was rounded up to 0.003 because `sc_stats` only provides three significant figures. The Mandarin ST system was judged significantly better than CL at a level of $p = 0.003$ (denoted *), and the Swahili system at a level of $p < 0.001$ (denoted **); the Cantonese and Hungarian ST systems were judged to be not significantly better than CL.

The column headed PT-ADAPT in Table VII lists PERs from ASR systems that have been adapted to PTs in the target language. The relative reductions in PER compared to both baselines are listed in the last two columns. Reductions on the evaluation set were tested for statistical significance using the MAPSSWE test of the `sc_stats` tool. All differences were found to be statistically significant at $p < 0.001$ (denoted **). This suggests that adaptation with PTs is providing more information than that obtained by model self-training alone. It is also interesting that PER improvements for Swahili are larger than for the other three languages. We conjecture this may be due to the relatively good mapping between Swahili's phone inventory and that of English. For example: all Swahili vowel qualities are also found in English, and the Swahili phonemes that would be unfamiliar to an English speaker (prenasalized stops, palatal consonants) have representations in English orthography that are fairly natural ("mb", "nd", etc. for prenasalized stops; "tya", "chya", "nya", etc. for palatals). In contrast: Mandarin, Cantonese, and Hungarian each have at least two vowel qualities not found in English; Mandarin and Cantonese have many diphthongs not found in English; and some of the consonant phonemes (e.g., Mandarin retroflexes) do not have representations in English orthography that are obvious or straightforward. Another possible source of the difference would be if Swahili phonemes had (on average) higher perceptual discriminability (to English speakers) than the phonemes of other languages tested. In other words, there may have been fewer many-to-one mappings in Swahili than in the other languages. Both of these factors (orthographic and perceptual) may have led to less label noise in the Swahili PT than in the PTs of the other languages.

It is also useful to compare the performance of GMM-HMM and NN-HMM systems. In the CL setting, an ASR trained using six languages is then applied to an unseen seventh language, without adaptation; in this setting, the NN consistently outperforms the GMM. In the PT-ADAPT setting, GMMs and NNs are adapted using PTs in the target language. PT adaptation improves the performance of both types of ASR, but the adapted NN does not consistently outperform the GMM across all tested languages.

VII. DISCUSSION

Models of human neural processing systems have often been used to inspire improvements in machine-learning systems (for a catalog of such approaches and a warning, see [6]). These systems are often called neuromorphic, because the system is engineered to mimic the behavior of human neural systems. In contrast to that approach, our incorporation of EEG signals into ASR resonates with the Human Aided Computing approach used in computer vision [46], [54]. Together with our EEG work presented here, this class of approach represents a less explored direction for design of machine learning systems, whereby recorded neural data (rather than neuro-inspired models) are used as a source of prior information to improve system performance. Therefore, our work here suggests that, by thinking about the kinds of prior information required by a machine learning system, engineers and neuroscientists can work together to design specific neuroscience experiments that leverage human abilities and provide information that can be directly integrated into the system to solve an engineering problem.

NN-HMM outperforms the GMM-HMM in all baseline conditions, but not always when adapted using PTs. Preliminary analysis suggests that the NN is more adversely affected than the GMM by label noise in the PTs. A NN is trained to match the senone posterior probabilities $\pi(s_t^\ell | x^\ell, \phi^\ell, \theta)$ computed by a first-pass GMM-HMM. Many papers have demonstrated that entropy in the senone posteriors is detrimental to NN training, and that the senone posteriors should therefore be quantized ($\pi(s_t^\ell) \rightarrow \{0, 1\}$) prior to NN training. In PT adaptation, however, entropy is unavoidable, and quantizing the forced alignment doesn't necessarily help. Fig. **??** showed that the 1-best path through the PT is only correct for 29-49% of all phones, depending on language. There is good reason for this: the transcribers don't speak the target language, so they find some of its phone pairs to be perceptually indistinguishable. Future work will seek methods that can improve the robustness of NN training in the face of label noise.

This paper has tentatively defined an "under-resourced language" to be one that lacks transcribed speech data. Other authors have proposed that if a language lacks transcribed speech, ASR can be initialized in that language by adapting a cross-lingual baseline. Other authors have proposed, and Table VII confirms, that significant error reductions can be achieved using self-training: by automatically labeling speech in the target language, and adding the self-labeled data to the training set. Table VII shows that further error rate reductions can be achieved using mismatched crowdsourcing: by asking non-speakers of the target language to write down what they hear, and by interpreting their nonsense orthography as information about the phonetic content of the utterances. The PER of mismatched crowdsourcing (Fig. **??**) is almost as high as the PER of cross-language ASR (Table VI), but the information provided by mismatched crowdsourcing is superior to that provided by self-training in the sense that it trains a better ASR.

VIII. Conclusions

When a language lacks transcribed speech, other types of information about the speech signal may be used to train ASR. This paper proposes compiling the available information into a probabilistic transcription: a pmf over possible phone transcriptions of each waveform. Three sources of information are discussed: self-training, mismatched crowdsourcing, and EEG distribution coding. Experiments demonstrate that self-training outperforms cross-lingual ASR in two of the four test languages (Mandarin and Swahili). Adaptation using mismatched crowdsourcing outperforms both cross-lingual ASR and self-training in all four of the test languages. Auxiliary information from EEG is used, together with text-based phone language models, to improve the decoding of transcripts from mismatched crowdsourcing.

IX. Acknowledgments

The work reported here was started at JSALT 2015 at UW, Seattle, and was supported by JHU via grants from NSF (IIS), DARPA (LORELEI), Google, Microsoft, Amazon, Mitsubishi Electric, and MERL. Parts of this work were previously published in [31].

REFERENCES

[1] Amazon. Mechanical Turk. http://www.mturk.com.

[2] R Baayen, R Piepenbrock, and L Gulikers. CELEX2. Technical Report LDC96L14, Linguistic Data Consortium, 1996.

[3] James Baker. The dragon system — an overview. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 23:24–29, 1975.

[4] Y. Bengio, R. De Mori, G. Flammia, and R. Kompe. Global optimization of a neural network - hidden Markov model hybrid. *IEEE Trans. Neural Networks*, 3(2):252–259, 1992.

[5] V. Berment. *Méthodes pour informatiser des language des langues et des groupes de langues peu dotées*. PhD thesis, J. Fourier University, Grenoble, 2004.

[6] Hervé Bourlard, Hynek Hermansky, and Nelson Morgan. Towards increasing speech recognition error rates. *Speech Communication*, 18:205–231, 1996.

[7] Alexandra Canavan and George Zipperlen. CALLFRIEND egyptian arabic. Technical Report LDC96S49, Linguistic Data Consortium, 1996.

[8] Özgür Cetin. Unsupervised adaptive speech technology for limited resource languages: A case study for Tamil. In *Workshop on Spoken Language Technology for Under-Resourced Languages (SLTU)*, Hanoi, Vietnam, 2008.

[9] George E. Dahl, Dong Yu, Li Deng, and Alex Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Trans. Audio, Speech and Language*, 20(1):30–42, 2012.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39:1–38, 1977.

[11] Giovanni di Liberto, James A. O'Sullivan, and Edmund C. Lalor. Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, 25(19):2457–2465, 2015.

[12] Van Hai Do, Xiong Xiao, Eng Siong Chng, and Haizhou Li. Context dependant phone mapping for cross-lingual acoustic modeling. In *International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 16–19, 2012.

[13] Mohamed Elmahdy, Mark Hasegawa-Johnson, and Eiman Mustafawi. Development of a TV broadcasts speech recognition system for Qatari Arabic. In *The 9th edition of the Language Resources and Evaluation Conference (LREC 2014)*, Reykjavik, Iceland, 2014.

[14] Jean-Luc Gauvain and Chin-Hui Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298, 1994.

[15] František Grézl, Martin Karafiaát, and Karel Veselý. Adaptation of multilingual stacked bottle-neck neural network structure for new language. In *Proc. ICASSP*, pages 7704–7708, 2014.

[16] Mark Hasegawa-Johnson. WS15 dictionary data. Downloaded 9/24/2015 from http://isle.illinois.edu/sst/data/dict, 2015.

[17] Hynek Hermansky, Dan Ellis, and Sangita Sharma. Tandem connectionist feature extraction for conversational HMM systems. In *Proc. ICASSP*, Istanbul, Turkey, 2000.

[18] Roger Hsiao, Tim Ng, František Grézl, Damianos Karakos, Stavros Tsakalidis, Long Nguyen, and Richard Schwartz. Discriminative semi-supervised training for keyword search in low resource languages. In *Proc. Interspeech*, pages 440–443, 2013.

[19] Jui-Ting Huang, Jing Li, Dong Yu, Li Deng, and Yifan Gong. Cross language knowledge transfer using multilingual deep neural network with shared hidden layers. In *Proc. ICASSP*, 2013.

[20] Po-Sen Huang and Mark Hasegawa-Johnson. Cross-dialectal data transferring for gaussian mixture model training in Arabic speech recognition. In *International Conference on Arabic Language Processing (CITALA)*, pages 119–122, 2012.

[21] Yan Huang, Dong Yu, Yifan Gong, and Chaojun Liu. Semi-supervised GMM and DNN acoustic model training with multi-system combination and re-calibration. In *Proc. Interspeech*, pages 2360–2363, 2013.

[22] David Imseng, Petr Motlicek, Hervé Bourlard, and Philip N. Garner. Using out-of-language data to improve an under-resourced speech recognizer. *Speech Communication*, 56:142 – 151, 2014.

[23] International Phonetic Association (IPA). International phonetic alphabet, 1993.

[24] R. Jakobson, G. Fant, and M. Halle. Preliminaries to speech analysis. Technical Report 13, MIT Acoustics Laboratory, 1952.

[25] Bing-Hwang Juang and Lawrence Rabiner. The segmental K-means algorithm for estimating parameters of hidden Markov models. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 38(9):1639–1641, 1990.

[26] Preethi Jyothi and Mark Hasegawa-Johnson. Acquiring speech transcriptions using mismatched crowdsourcing. In *Proceedings of AAAI*, 2015.

[27] Preethi Jyothi and Mark Hasegawa-Johnson. Transcribing continuous speech using mismatched crowdsourcing. In *Proceedings of Interspeech*, 2015.

[28] Yamuna Kachru. Hindi-Urdu. In Bernard Comrie, editor, *The world's major languages*, pages 399–416. Oxford University Press, New York, 1990.

[29] S. Krauwer. The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. In *Proc. International Workshop Speech and Coputer (SPECOM-2003)*, pages 8–15, Moscow, Russia, 2003.

[30] Kevin Lenzo. The CMU pronouncing dictionary. Downloaded 1/31/2016 from http://www.speech.cs.cmu.edu/cgi-bin/cmudict, 1995.

[31] Chunxi Liu, Preethi Jyothi, Hao Tang, Vimal Manohar, Rose Sloan, Tyler Kekona, Mark Hasegawa-Johnson, and Sanjeev Khudanpur. Adapting ASR for under-resourced languages using mismatched transcriptions. In *Proc. ICASSP*, 2016.

[32] Jonas Lööf, Christian Gollan, and Hermann Ney. Cross-language bootstrapping for unsupervised acoustic model training: rapid development of a Polish speech recognition system. In *Proceedings of Interspeech*, 2009.

[33] L. Mangu, E. Brill, and A. Stolcke. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer, Speech and Language*, 14(4):373–400, 2000.

[34] Aanchan Mohan, Richard Rose, Sina Hamidi Ghalehjegh, and S. Umesh. Acoustic modelling for speech recognition in Indian languages in an agricultural commodities task domain. *Speech Communication*, 56:167 –180, 2014.

[35] Mehryar Mohri, Fernando Pereira, and Michael Riley. Weighted finite-state transducers in speech recognition. *Computer Speech and Language*, 16:69–88, 2002.

[36] Mehryar Mohri, Fernando Pereira, and Michael Riley. Speech recognition with weighted finite-state transducers. In *Springer Handbook of Speech Processing*, pages 559–584. Springer, 2008.

[37] Steven Moran, Daniel R. McCloy, and Richard A. Wright. PHOIBLE: Phonetics Information Base and Lexicon Online, 2013.

[38] Nelson Morgan and Hervé Bourlard. Continuous speech recognition. *IEEE Signal Processing Magazine*, 12(5):25–42, 1995.

[39] D.S. Pallet, W.M. Fisher, and J.G. Fiscus. Tools for the analysis of benchmark speech recognition tests. In *Proc. ICASSP*, volume 1, pages 97–100, 1990.

[40] D. Povey, A. Ghoshal, et al. The Kaldi speech recognition toolkit. *Proc. of ASRU*, 2011.

[41] Daniel Povey, Lukáš Burget, Mohit Agarwal, Pinar Akyazi, Feng Kai, Arnab Ghoshal, Ondrej Glembek, Nagendra Goel, Martin Karafiát, Ariya Rastrow, Richard C. Rose, Petr Schwarz, and Samuel Thomas. The subspace Gaussian mixture model—a structured model for speech recognition. *Computer Speech and Language*, 25(2):404–439, 2011.

[42] S. Scanzio, P. Laface, L. Fissore, R. Gemello, and F. Mana. On the use of a multilingual neural network front-end. In *Proc. Interspeech*, pages 2711–2714, 2008.

[43] Tanja Schultz and Alex Waibel. Experiments on cross-language acoustic modeling. In *INTERSPEECH*, 2001.

[44] H.J. Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Trans. on Information Theory*, 11:363–371, 1965.

[45] Claude Shannon and William Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, 1949.

[46] Pradeep Shenoy and Desney S. Tan. Human-aided computing: utilizing implicit human processing to classify images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 845–854, Florence, 2008. ACM Press.

[47] Ke Chai Sim and Haizhou Li. Context sensitive probabilistic phone mapping model for cross-lingual speech recognition. In *Proc. Interspeech*, pages 2175–2718, 2008.

[48] Special Broadcasting Services Australia. Podcasts in Your Language. http://www.sbs.com.au/podcasts/yourlanguage.

[49] A. Stolcke, F. Grezl, M.-Y. Hwang, X. Lei, N. Morgan, and D. Vergyri. Cross-domain and cross-lingual portability of acoustic features estimated by multilayer perceptrons. In *Proc. ICASSP*, 2006.

[50] P. Swietojanski, A. Ghoshal, and S. Renals. Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR. In *Proc. IEEE Workshop on Spoken Language Technology (SLT)*, 2012.

[51] Karel Vesely, Mirko Hannemann, and Lukas Burget. Semi-supervised training of Deep Neural Networks. In *Proceedings of ASRU*, 2013.

[52] Karel Vesely, Martin Karafiát, Frantisek Grezl, Marcel Janda, and Ekaterina Egorova. The language-independent bottleneck features. In *Proceedings of SLT*, 2012.

[53] Ngoc Thang Vu, Franziska Kraus, and Tanja Schultz. Cross-language bootstrapping based on completely unsupervised training using multilingual A-stabil. In *Proceedings of ICASSP*, 2011.

[54] Jun Wang, Eric Pohlmeyer, Barbara Hanna, Yu-Gang Jiang, Paul Sajda, and Shih-Fu Chang. Brain state decoding for rapid image retrieval. In *Proceedings of the 17th ACM international conference on multimedia*, pages 945–954. ACM Press, 2009.