Low-Resource Spoken Keyword Search Strategies in Georgian Inspired by Distinctive Feature Theory

Nancy F. Chen[1], *Boon Pang Lim*[1], *Van Hai Do*[3], *Van Tung Pham*[2], *Chongjia Ni*[1], *Haihua Xu*[2], *Mark Hasegawa-Johnson*[3,4], *Wenda Chen*[4], *Xiong Xiao*[2], *Sunil Sivadas*[1], *Eng Siong Chng*[2], *Bin Ma*[1], *Haizhou Li*[1]

[1]Institute for Infocomm Research, Singapore; [2]Nanyang Technological University, Singapore; [3]Advanced Digital Sciences Center, Singapore; [4]University of Illinois Urbana-Champaign, USA

nfychen@i2r.a-star.edu.sg

Abstract

We present low-resource spoken keyword search (KWS) strategies guided by distinctive feature theory in linguistics to conduct data selection, feature selection, and transcription augmentation. These strategies were employed in the context of the 2016 NIST Open Keyword Search Evaluation (OpenKWS16) using conversational Georgian from the IARPA Babel program. In particular, we elaborate on the following: (1) We exploit glottal-source-related acoustic features that characterize Georgian ejective phonemes (`[+constricted glottis]`, `[+raised larynx ejective]` specified in distinctive feature theory). These features complement standard acoustic features, leading to a relative fusion gain of 11.9%. (2) We use noisy channel models to incorporate probabilistic phonetic transcriptions from mismatched crowdsourcing to conduct transfer learning to improve KWS for extremely under-resourced conditions (24 min of transcribed Georgian), achieving a relative improvement of 118% over the baseline and a relative fusion gain of 32%.(3) Using distinctive feature analysis, we select a compact subset of source languages used in past evaluations to ensure high phonetic coverage for cross-lingual acoustic modeling when only limited system development time and computational resources are available. This strategy leads to comparable performance to using all available linguistic resources when only 1/3 of the source languages were chosen.

Index Terms: Spoken term detection (STD), keyword spotting, multilingual training, automatic speech recognition (ASR)

1. Introduction

Spoken keyword search (KWS) can be cast as a detection or a retrieval task, where the objective is to find all occurrences of an orthographic term (be it word or phrase) from large streams of audio recordings. Approaches to spoken KWS are often based on large vocabulary continuous speech recognition (LVCSR), following the transcribe-and-search paradigm. For resource-rich languages such as Mandarin, Arabic, or English, high performance is readily achieved via copious amounts of transcribed audio [1]. However, low-resource languages such as Georgian are more challenging due to the lack of word-transcribed training data. Such challenges have led to initiatives such as the NIST Open Keyword Search Evaluation and the IARPA Babel program: "... to rapidly develop speech recognition capability for keyword search in a previously unstudied language, with limited amounts of transcription."

Researchers commonly address challenges of low-resource spoken KWS via two avenues. The first avenue bypasses the problem by improving the overall performance of spoken keyword search. Such approaches include feature extraction and selection [2], keyword verification or rescoring [3, 4], score normalization [5, 6], and system combination (fusion) [5, 6]. This approach is popular in a time constrained evaluation setup, because once a baseline LVCSR-based KWS system is set up, by altering the system input (acoustic features) and/or output (posterior scores), one can efficiently obtain system combination gains effectively. The second type of approaches for low-resource spoken KWS tackles the data sparsity problem directly at various levels such as data augmentation [7], data selection [8], data-efficient training [9, 10], or linguistically-inspired frameworks [11]. In this work, we focus on the latter avenue by exploiting distinctive feature theory in linguistics [12, 13] in proposing economical strategies in data selection for cross-lingual transfer learning in acoustic modeling, feature selection, and mismatched crowdsourcing [14]. When there are massive amounts of target training data and computational resources, such linguistic guidance might be unnecessary. Yet, in the absence of such resources, linguistically guided approaches could be fruitful. A system diagram of our proposed KWS system inspired by distinctive feature analysis is in Figure 1.

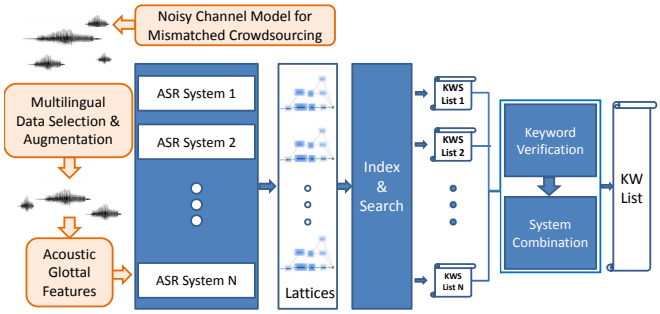

Figure 1: *SINGA spoken keyword search system for low-resource languages. Blocks filled with light orange are inspired by distinctive feature theory: Multilingual Data Selection and Augmentation, Glottal Acoustic Feature Extraction, and Noisy Channel Model for Mimsmatched Crowdsourcing. Other low-resource strategies are elaborated in [8, 15, 16].*

2. Relation to prior work

There are more than 6,700 spoken languages in the world, all of them evolved to be produced and perceived by humans. All languages therefore exploit nonlinearities of human articulation and perception to increase the *distinctiveness* of words, and to

organize morphophonemic variation. A phonological distinction used by a higher-than-chance fraction of languages is called a *distinctive feature* [12, 13]; linguists generally agree that all distinctions in all languages in the world can be encoded by $\sim$ 40 distinctive features. A distinctive feature is either an acoustic attribute or an articulatory gesture. Binary values signify whether a phonetic segment is specified by a distinctive feature: a positive value, `[+]`, denotes the presence of the feature, while a negative value, `[-]`, denotes the absence of the feature. Distinctive features are grouped into different categories: major class features such as `[+/-sonorant]`; laryngeal features such as `[+/-voicing]`; manner features such as `[+/-nasal]`; place features `[+/-labial]`; and vowel space features such as `[+/-back]`.

Distinctive features `[+constricted glottis]` and `[+raised larynx ejective]` are used in an unusual way in Georgian. In this work we investigate the usefulness of such glottal related acoustic features in spoken KWS on conversational telephone speech of Georgian in Section 4 and related prior work in Section 2.1.

Distinctive feature theory suggests that for under-resourced languages with few native speakers to perform transcription tasks, we can leverage on non-speakers of the language to infer ground-truth phonetic transcriptions (mismatched crowdsourcing) [14] since mutually unintelligible languages are characterized with overlapping articulatory gestures and acoustic properties. We discuss how prior work in mismatched crowdsourcing is extended in Section 2.2.

In the absence of sufficient system development time, computational power, or linguistic resources, distinctive feature theory can guide us to select a compact set of source languages to train a multilingual acoustic model to be readily fine-tuned for a new target language. We discuss related prior work on data selection and language selection in Section 2.3.

2.1. Acoustic Features Related to Glottal Source

We examine two glottal related distinctive features that are used in an unusual way in Georgian: `[+constricted glottis]` and `[+raised larynx ejective]`, and discuss acoustic features that can help characterize such speech production modes. `[+constricted glottis]` specifies glottal constriction, which leads to *creaky voice quality*. (Related terms: *glottalization, irregular phonation, and vocal fry.*) `[+constricted glottis]` is peculiar to certain languages: the *broken* Vietnamese toneme (ng$\tilde{a}$), the Tagalog glottalization phoneme, and Georgian ejective stop consonants [17]. Glottalization also occurs in American English in phrases such as "uh-oh" [18]. `[+raised larynx ejective]` refers to simultaneous constriction in the oral cavity and glottis, which often results in loud bursts from increased oral air pressure from glottal constriction [17]. In Georgian, ejection is only heard at word-initial positions, but *glottalization* can be spread through co-aritculation to the following vowel and to word medial positions [17]. The most common source-related acoustic feature is the pitch estimate F0, which is known to improve ASR in both tonal and non-tonal languages. The glottalization phoneme in Tagalog has been modeled by exploring different hidden Markov model topologies and using voicing features [19]. Fundamental frequency variation (FFV) has also shown to improve ASR in both tonal and non-tonal languages [20], and in identifying tonal mispronunciation in Mandarin [21]. Creaky voice quality (CVQ) features are less explored in speech technology, though they have been used to model English allophones [22]. In this work, we examine how CVQ features work on conversational Georgian in spoken keyword search tasks.

2.2. Noisy Channel Models for Mismatched Crowdsourcing

Obtaining ground-truth labels is essential in supervised machine learning. Crowdsourcing is a cost-effective way to obtain ground-truth labels for many machine learning tasks in spoken language technology, where the human transcribers are usually native speakers of the target language (e.g., Georgian). Mismatched crowdsourcing asks non-speakers (e.g., Mandarin) of the target language to write what they hear, and their nonsense transcripts (*mismatched transcriptions*) are decoded using noisy channel models of second language speech perception [14]. These *mismatched transcriptions* tend to correctly transcribe distinctive features shared by both the target language (Georgian) and the annotation language (Mandarin), but features foreign to the annotator tend to be incorrectly transcribed. Recent work from the 2nd Frederick Jelinek Memorial Summer Workshop [23] used such mismatched transcriptions to train acoustic models for improved phone error rate (PER). In this work, we extend mismatched crowdsourcing from PER tasks using podcasts (semi-broadcast news speaking style over clean channels) to spoken keyword search tasks in noisy conversational telephone speech in extremely under-resourced scenarios.

2.3. Data Selection for Cross-Lingual Acoustic Modeling

Data selection approaches are often used in active learning, where the goal is to find the most informative and representative subset of audio for human transcription. These approaches include utility scores (e.g., confidence scores from acoustic models) [24, 25], entropy computation [26, 27, 28, 29], or submodular optimization [30, 31, 32, 8]. The aforementioned work primarily focuses on selecting data from a larger corpus of the same language. When considering data from source languages different from the target language, language identification (LID) has helped identify languages to train cross-lingual acoustic models when the target language is known [33]. In the context of OpenKWS16, systems had to be developed within one week after the release of data, making the LID approach infeasible due to the short development time. Minimal system development time is essential in providing situational awareness information from any language in emergent missions such as humanitarian assistance[1], as in the case of the catastrophic 2010 Haiti earthquake. To tackle such computational challenges, we need to select linguistic resources that are comprehensive in terms of acoustic phonetic coverage so we can fine-tune it swiftly to the target language in urgent situations. In this work, we turn to an approach grounded in linguistic theory. Distinctive features have been used to estimate the phonetic coverage and information loss of mismatched crowdsourcing, where human transcribers are asked to annotate a language they do not speak [34]. Sharing a similar spirit, we use distinctive features to analyze the phonetic coverage of existing transcribed multilingual resources used for cross-lingual transfer in acoustic modeling.

3. Experimental Setup

3.1. Corpora

This effort uses the Georgian language release (IARPA-babel404b-v1.0a) for the NIST OpenKWS16 Evaluation. The Full Language Pack (FLP) training set includes 40 hrs of conversational telephone speech. The Very Limited Language Pack

[1]DARPA LORELEI: http://www.darpa.mil/program/low-resource-languages-for-emergent-incidents

Table 1: *Glottal Feature Experiment*

System	Features	ATWV
G0	PLP+pitch+BNF	0.3107
G1	PLP+pitch+BNF+FFV	0.3235
G2	PLP+pitch+BNF+CVQ	0.3123
G1+G2	(PLP+pitch+BNF+FFV) + (PLP+pitch+BNF+CVQ)	0.3610

Table 2: *Mismatched Crowdsourcing Experiment*

System	Description	ATWV	MTWV
C1	24-min transcribed Georgian audio	0.0317	0.0619
C2	C1 adapted with 10-hr mismatched Mandarin transcription	0.0690	0.1036
C3	C1 + C2	0.0911	0.1231

(VLLP) provides word transcriptions for a 3-hr subset of FLP, a 3-hr tuning set, and a 10-hr developmental set. All results reported are on the 10-hr development data. Pronunciation lexicons were not offered, but web data was. OpenKWS16 allowed 24 Babel languages and some LDC data[2] for acoustic modeling.

3.2. Evaluation Metric

Term-weighted value (TWV) is 1 minus the weighted sum of miss $P_{\text{miss}}(\theta)$ and false alarm $P_{\text{FA}}(\theta)$: $\text{TWV}(\theta) = 1 - [P_{\text{miss}}(\theta) + \beta P_{\text{FA}}(\theta)]$; θ is the decision threshold. Actual term-weighted value (ATWV) is TWV of the chosen threshold; maximum term-weighted value (MTWV) is the best TWV of all θ. $\beta = 0.99999$ for NIST OpenKWS16, thus penalizing miss probability heavily.

4. Glottal Feature Experiment

4.1. Acoustic Features Beyond Pitch

4.1.1. Fundamental Frequency Variation (FFV) Features

In contrast to scalar representations of pitch, FFV [35] is a vector representation of delta pitch. FFV is obtained using two asymmetric windows placed on the same frame, one emphasizing earlier samples and the other emphasizing later samples. Their corresponding spectra are compared using different assumptions for the rate of pitch variation to infer delta pitch. This comparison results in a spectrum, which is further reduced to a 7-dim vector for each speech frame by applying 7 trapezoidal filters centered around different rates of frequency variation.

4.1.2. Creaky Voice Quality (CVQ) Features

Creaky voice (*glottalization*) is caused by glottal constriction. Features such as pitch or FFV, while related to the glottis, are not explicitly designed to characterize the irregular phonations from strong glottal constriction. Therefore, we adopt two additional features: (1) Amplitude difference of the 1st and 2nd harmonics of the inverse-filtered voice signal (H1*-H2*), which is the acoustic correlate of the open quotient of the glottis [36]. (2) Mean autocorrelation ratio, a temporal measure more robust to channel effects [22]. CVQ features were extracted in voiced regions; unvoiced regions are padded with zero [22].

4.2. System Implementation Details

The acoustic models were progressively trained, starting from GMM systems, going from monophones to triphones, applying LDA and MLLT, eventually arriving at a final GMM system with speaker adaptive training using feature MLLR. These GMMs were used to generate alignments which were used for hybrid DNN system training. This 6-layer feed forward DNN was used to generate a new set of alignments, used for bottleneck feature (BNF) training. The features were passed through the bottleneck network and used to train a second bottleneck network, resulting in stacked BNFs [2]. We retrained our hybrid-DNN using these stacked BNFs, and used sequence discriminative (sMBR) training to further refine it. A 4-gram K-N smoothed language model was trained using text transcripts augmented with sentences filtered from the official web data provided by BBN.

[2]See OpenKWS16 Evaluation Plan (https://www.nist.gov/document-194) & OpenKWS16LDC Data Agreement (https://www.nist.gov/sites/default/files/documents/itl/iad/mig/OpenKWS16_LDCData_EvalAgreement-V1_LDCRev.pdf).

Table 3: *24 Babel languages and word error rates of their corresponding monolingual GMM systems (FLP condition).*

Language Family Category	Language	WER (%)
Southeast Asian & Tonal	Cantonese	65.1
	Lao	58.1
	Vietnamese	61.6
Southeast Asian Austronesian	Cenuano	65.7
	Javanese	73.2
	Tagalog	62.2
	Tok Pisin	51.0
South Asian (Indian)	Assamese	68.5
	Bengali	68.9
	Tamil	75.1
	Telugu	79.1
Middle Eastern	Kazakh	66.3
	Kurdish	78.1
	Mongolian	68.9
	Pashto	62.2
	Turkish	64.4
African and Caribbean	Amharic	62.7
	Dholuo	58.4
	Haitian Creole	61.1
	Igbo	73.5
	Swahili	57.6
	Zulu	72.0
Native American	Guarani	62.9
Eastern European	Lithuanian	61.9

4.3. Results

Table 1 shows that appending FFV (System G1 to the baseline setup of PLP+pitch+BNF (System G0) leads to 4.1% relative gain, while appending CVQ features (System G2) results in only 0.5% relative gain. However, when we combine system G1 and system G2, we achieve 11.9 % relative gain compared to System G1, suggesting that FFV and CVQ are complementary.

5. Mismatched Crowdsourcing Experiment

5.1. Baseline System C1

Mismatched crowdsourcing is suitable for extremely under-resourced scenarios, where native transcriptions are minimal. We use Georgian as a test case in this paper. The baseline KWS system is trained on 24 minutes of Georgian audio randomly extracted from the 3-hr VLLP dataset. The acoustic features are MFCC and pitch, used to train a DNN with 4 hidden layers, each with 1024 nodes. The number of senone outputs are 364 (number of tied triphones). A 4-gram K-N smoothed language model was used for decoding.

Table 4: *Language Selection Experiment*

System	# of Languages	#hr	ATWV
UpperBound S0	27 (24 Babel languages + 3 from LDC: Arabic, Mandarin, Spanish)	1,954	0.6498
Proposed S1	10 from Babel: Bengali, Haitian Creole, Lao, Kurdish, Zulu, Kazakh, Lithuanian, Guarani, Amharic, Javanese	480	0.6296
Proposed S2	9 from Babel: Cantonese, Assamese, Pashto, Turkish, Tagalog, Zulu, Lithuanian, Guarani, Amharic	615	0.6383

5.2. Proposed System C2

We adapted the baseline system C1 with mismatched transcriptions: 10 hrs from the untranscribed portions of FLP were chosen by maximizing the number of speakers via choosing 70 seconds from the middle of each untranscribed Georgian conversation. A total of 8 Mandarin transcribers were hired from Upwork (https://www.upwork.com), each in charge of 2.5 hrs, so each Georgian audio file was transcribed by 2 Mandarin speakers using Pinyin (romanized phonetic system). In this work, mismatched transcription is converted to matched transcription using a mismatched channel, modeled as a finite memory process using weighted finite state transducers (WFST) [14, 23]. The weights on the arcs of the WFST model are learned using the EM algorithm[3] to maximize the likelihood of the observed training instances. OpenFST [37] is used for all finite-state operations. Mandarin phones are decoded via the mismatched channel into Georgian phones, in a lattice format called *probabilistic transcription (PT)* [14, 23]. PT is used to adapt the baseline system C1. Unlike previous studies [38, 23], where no native transcription was used, we assume there is a limited amount (24 minutes) of native Georgian transcriptions to train an initial acoustic model (System C1), and 10 hrs of PT and its corresponding audio are used to further adapt C1.

5.3. Results and Discussion

Table 2 shows that the proposed system (C2) improves the baseline (C1) by 118% relative in ATWV, and combining the two systems further improves performance by 32% relative to the proposed system C2. In addition to ATWV, we also show MTWV results since the difference between ATWV and MTWV suggest further gains can be obtained by more targeted calibration and normalization. These investigations are test beds for on-going work on ASR for under-resourced languages such as Singapore Hokkien (Min Nan), where native transcriptions are extremely challenging to acquire given the absence of a formal writing system. We are also examining how to resolve the noisy label problem of mismatched transcriptions.

6. Language Selection Experiment

6.1. Oracle System (Upper Bound S0)

A total of 43-dimension filter bank features were extracted from all 24 languages in the Babel data and 3 languages in the LDC set (Arabic, Mandarin, Spanish). The shared-hidden-layer multilingual DNN consists of 6 shared hidden layers (each with 2048 nodes); the output softmax layer is fine-tuned using 40 hrs of the FLP Georgian data. For pronunciation modeling, 1-letter graphemes were used to approximate phonemes. A 3-gram LM was estimated using NIST provided web data.

6.2. Proposed Systems

Our objectives in selecting a subset of languages are as follow:

(1) Maximize acoustic phonetic diversity. For each of the 24 Babel languages we compared their distinctive features, similar to [34], using phonological inventory data (http://phoible.org). Zulu and Amharic are two of the most comprehensive languages that cover 29 and 28 distinctive features respectively. The union of Zulu and Amharic reach 33 distinctive features among 38 specified distinctive features. The union of the distinctive features in Zulu and Amharic excluding the 22 common distinctive features across all languages: `+anterior, +click, +constricted glottis, +long, +lowered larynx implosive, +raised larynx ejective, +spread glottis, +tap, +tense, +tone, +trill`. For distinctive features not in this union set of cardinality of 33, they are not specified in the other Babel languages either. Zulu and Ahmaric were always chosen to ensure we include not only common phonological features, but also as many peculiarities as possible.

(2) Minimize number of languages. Besides taking a phonetic perspective in language selection through distinctive feature analysis, we also considered historical linguistic lineage and geographical proximity to group the 24 Babel languages into language family categories shown in Table 3. Each category shares similarities beyond the acoustic-phonetic level, including phonological structure and prosodic rhythm; e.g., the Southeast Asian Tonal category are all monosyllabic tonal languages. Note that the categorization is more refined for Asian languages due to its large proportion of 50%.

(3) Exclude anomalies. High word error rate implies potential issues to exclude the language in multilingual training: low signal-to-noise ratio, cross-talk, poor transcription quality, and linguistic peculiarities such as agglutination, which could be effectively modeled using automatically parsed morphs [8].

The experimental setup is the same as the oracle system (S0 in Table 4) except the languages used to train the shared-hidden layer multilingual DNNs (SHL-MDNN) are different. See Table 4 for details. Table 4 suggests that our language selection strategy only sacrifices less than 2% relative ATWV, when choosing only 1/3 of the total languages.

7. Conclusion

We presented low-resource spoken keyword search strategies guided by distinctive feature theory to conduct acoustic feature selection, ground-truth transcription augmentation, and data/language selection: (1) We exploited glottal features that characterize Georgian ejective phonemes and showed that they complement standard acoustic features, leading to fusion gains. (2) We used noisy channel models of second language speech perception to incorporate probabilistic phonetic transcriptions from mismatched crowdsourcers to improve keyword search performance for extremely low-resource conditions. (3) Using distinctive feature analysis coupled with linguistic lineage, we selected a compact subset of source languages to ensure high phonetic coverage for cross-lingual acoustic modeling.

[3] Carmel finite-state toolkit," http://www.isi.edu/licensed-sw/carmel

8. References

[1] J. G. Fiscus, J. Ajot, J. S. Garofolo, and G. Doddingtion, "Results of the 2006 spoken term detection evaluation," in *ACM SIGIR Workshop on Searching Spontaneous Conversational*, 2007, pp. 51–55.

[2] F. Grézl, M. Karafiát, and K. Vesely, "Adaptation of multilingual stacked bottle-neck neural network structure for new language," in *Proc. of IEEE ICASSP*, 2014, pp. 7654–7658.

[3] V. Soto, L. Mangu, A. Rosenberg, and J. Hirschberg, "A comparison of multiple methods for rescoring keyword search lists for low resource languages," *Interspeech*, 2014.

[4] H.-y. Lee, Y. Zhang, E. Chuangsuwanich, and J. Glass, "Graph-based re-ranking using acoustic feature similarity between search results for spoken term detection on low-resource languages," *Interspeech*, pp. 2479–2483, 2014.

[5] J. Mamou, J. Cui, X. Cui, M. J. Gales, B. Kingsbury, K. Knill, L. Mangu, D. Nolden, M. Picheny, B. Ramabhadran, R. Schluter, A. Sethy, and P. C. Woodland, "System combination and score normalization for spoken term detection," in *ICASSP*, 2013, pp. 8272–8276.

[6] D. Karakos, R. Schwartz, S. Tsakalidis, L. Zhang, S. Ranjan, T. T. Ng, R. Hsiao, G. Saikumar, I. Bulyko, L. Nguyen *et al.*, "Score normalization and system combination for improved keyword spotting," in *Proc. of IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop*, 2013, pp. 210–215.

[7] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," in *ICASSP*, 2014, pp. 5582–5586.

[8] N. F. Chen, C. Ni, I.-F. Chen, S. Sivadas, V. T. Pham, H. Xu, X. Xiao, T. S. Lau, S. J. Leow, B. P. Lim *et al.*, "Low-Resource Keyword Search Strategies for Tamil," in *Proc. ICASSP*, 2015, pp. 5366–5370.

[9] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafiát, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas, "Subspace Gaussian mixture models for speech recognition," in *Proc. of IEEE ICASSP*, 2010, pp. 4330–4333.

[10] I.-F. Chen, N. F. Chen, and C.-H. Lee, "A Keyword-Boosted sMBR Criterion to Enhance Keyword Search Performance in Deep Neural Network Based Acoustic Modeling," in *INTERSPEECH*, 2014.

[11] H. G. Ngo, N. F. Chen, B. M. Nguyen, B. Ma, and H. Li, "A Minimal-Resource Transliteration Framework for Vietnamese," in *Proc. of Interspeech*, 2014.

[12] R. Jakobson, G. Fant, and M. Halle, "Preliminaries to speech analysis: The distinctive features and their correlates. acoustics laboratory," MIT, Technical Report, Tech. Rep., 1952.

[13] K. N. Stevens, *Acoustic phonetics*. MIT press, 2000, vol. 30.

[14] P. Jyothi and M. Hasegawa-Johnson, "Acquiring speech transcriptions using mismatched crowdsourcing." in *AAAI*, 2015, pp. 1263–1269.

[15] N. F. Chen, S. Sivadas, B. P. Lim, H. G. Ngo, H. Xu, B. Ma, and H. Li, "Strategies for Vietnamese keyword search," in *Proc. IEEE ICASSP*, 2014, pp. 4121–4125.

[16] N. F. Chen, H. Xu, X. Xiao, C. Ni, I.-F. Chen, S. Sivadas, C.-H. Lee, E. S. Chng, B. Ma, and H. Li, "Exemplar-inspired strategies for low-resource spoken keyword search in Swahili," in *Proc. of ICASSP*, 2016, pp. 6040–6044.

[17] C. Vicenik, "An acoustic study of Georgian stop consonants," *Journal of the International Phonetic Association*, vol. 40, no. 01, pp. 59–92, 2010.

[18] L. Dilley, S. Shattuck-Hufnagel, and M. Ostendorf, "Glottalization of word-initial vowels as a function of prosodic structure," *Journal of Phonetics*, vol. 24, no. 4, pp. 423–444, 1996.

[19] K. Riedhammer, V. H. Do, and J. Hieronymus, "A Study on LVCSR and Keyword Search for Tagalog." Proc. of Interspeech, 2013.

[20] F. Metze, Z. A. W. Sheikh, A. Waibel, J. Gehring, K. Kilgour, Q. B. Nguyen, and V. H. Nguyen, "Models of tone for tonal and non-tonal languages," in *Proc. IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop*, 2013.

[21] R. Tong, N. F. Chen, B. P. Lim, B. Ma, and H. Li, "Tokenizing fundamental frequency variation for Mandarin tone error detection," in *Proc. of IEEE ICASSP*, 2015, pp. 5361–5365.

[22] T.-J. Yoon, X. Zhuang, J. Cole, and M. Hasegawa-Johnson, "Voice quality dependent speech recognition," in *Int. Symposium on Linguistic Patterns in Spontaneous Speech*, 2008.

[23] M. A. Hasegawa-Johnson, P. Jyothi, D. McCloy, M. Mirbagheri, G. M. di Liberto, A. Das, B. Ekin, C. Liu, V. Manohar, H. Tang, E. C. Lalor, N. F. Chen, P. Hager, T. Kekona, R. Sloan, and A. K. C. Lee, "ASR for Under-Resourced Languages From Probabilistic Transcription," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 25, no. 1, pp. 50–63, 2017.

[24] D. Hakkani-Tur, G. Riccardi, and A. Gorin, "Active learning for automatic speech recognition," in *Proc. of IEEE ICASSP*, 2002.

[25] L. Lamel, J.-L. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech & Language*, vol. 16, no. 1, pp. 115–129, 2002.

[26] D. Yu, B. Varadarajan, L. Deng, and A. Acero, "Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion," *Computer Speech & Language*, vol. 24, no. 3, pp. 433–444, 2010.

[27] N. Itoh, T. N. Sainath, D. N. Jiang, J. Zhou, and B. Ramabhadran, "N-best entropy based data selection for acoustic modeling," in *Proc. IEEE ICASSP*, 2012, pp. 4133–4136.

[28] T. Fraga-Silva, J.-L. Gauvain, L. Lamel, A. Laurent, V. B. Le, and A. Messaoudi, "Active learning based data selection for limited resource STT and KWS," in *Proc. of Interspeech*, 2015, pp. 3159–3163.

[29] S. Thomas, K. Audhkhasi, J. Cui, B. Kingsbury, and B. Ramabhadran, "Multilingual data selection for low resource speech recognition," in *Interspeech*, 2016, pp. 3853–3857.

[30] Y. Wu, R. Zhang, and A. Rudnicky, "Data selection for speech recognition," in *Proc. IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop*, 2007, pp. 562–565.

[31] H. Lin and J. Bilmes, "How to select a good training-data subset for transcription: Submodular active selection for sequences," in *INTERSPEECH*, 2009.

[32] K. Wei, Y. Liu, K. Kirchhoff, and J. Bilmes, "Using document summarization techniques for speech data subset selection." in *Proc. of HLT-NAACL*, 2013, pp. 721–726.

[33] Y. Zhang, E. Chuangsuwanich, and J. Glass, "Language ID-based training of multilingual stacked bottleneck features," in *Proc. Interspeech*, 2014, pp. 1–5.

[34] L. R. Varshney, P. Jyothi, and M. Hasegawa-Johnson, "Language coverage for mismatched crowdsourcing," in *Proc. of Information Theory and Applications Workshop*, 2016.

[35] K. Laskowski and Q. Jin, "Modeling instantaneous intonation for speaker identification using the fundamental frequency variation spectrum," in *ICASSP*, 2009, pp. 4541–4544.

[36] H. M. Hanson, "Glottal characteristics of female speakers: Acoustic correlates," *Journal of Acoustical Soceity of America*, vol. 101, p. 466, 1997.

[37] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri, "OpenFst: A general and efficient weighted finite-state transducer library," in *International Conference on Implementation and Application of Automata*. Springer, 2007, pp. 11–23.

[38] A. Das and M. Hasegawa-Johnson, "An investigation on training deep neural networks using probabilistic transcriptions," in *Proc. of Interspeech*, 2016, pp. 3858–3862.