# STRATEGIES FOR VIETNAMESE KEYWORD SEARCH

*Nancy F. Chen*[1], *Sunil Sivadas*[1], *Boon Pang Lim*[1], *Hoang Gia Ngo*[2],
*Haihua Xu*[3], *Van Tung Pham*[3], *Bin Ma*[1], *Haizhou Li*[1]

[1]Institute for Infocomm Research, Singapore, [2]National University of Singapore, Singapore,
[3]Nanyang Technological University, Singapore

*nfychen@i2r.a-star.edu.sg*

## ABSTRACT

We propose strategies for a state-of-the-art Vietnamese keyword search (KWS) system developed at the Institute for Infocomm Research ($I^2R$). The KWS system exploits acoustic features characterizing creaky voice quality peculiar to lexical tones in Vietnamese, a minimal-resource transliteration framework to alleviate out-of-vocabulary issues from foreign loan words, and a proposed system combination scheme FusionX. We show that the proposed creaky voice quality features complement pitch-related features, reaching fusion gains of 17.7% relative (6.9% absolute). To the best of our knowledge, the proposed transliteration framework is the first reported rule-based system for Vietnamese; it outperforms statistical-approach baselines up to 14.93 - 36.73% relative on foreign loan word search tasks. Using FusionX to combine 3 sub-systems, the actual term-weighted value (ATWV) reaches 0.4742, exceeding the ATWV=0.3 benchmark for IARPA Babel participants in the NIST OpenKWS13 Evaluation.

***Index Terms***— spoken term detection, audio indexing, large vocabulary continuous speech recognition (LVCSR), glottalization, deep neural networks (DNN), low-resourced languages

## 1. INTRODUCTION

With the large emphasis on big data and the grown amount of audio data collections, keyword search and retrieval is becoming increasingly important for military, intelligence and civilian applications. Keyword search is a detection task where the goal is to find all occurrences of an orthographic term (e.g., word or phrase) from audio recordings. In contrast to previous work on keyword spotting where customized detectors are built for pre-specified query terms (e.g., [1, 2]), pre-indexed keyword search systems are designed without knowledge of the query terms. These types of keyword search systems are more flexible in handling the increase in the number of keywords and the absence of keywords in the training data. Automatic speech recognition (ASR) is often an essential component for such pre-indexed keyword search systems.

ASR paradigms for keyword search have worked well on resource-rich languages like English, as have been shown in the NIST 2006 Spoken Term Detection Evaluation [3]. However, such *transcribe-and-search* approaches pose particular challenges to under-resourced languages such as Vietnamese, as ASR typically rely on copious amounts of transcribed speech to achieve high performance. To tackle these challenges, the IARPA Babel program aims to foster research "to rapidly develop speech recognition capability for keyword search in a previously unstudied language, working with speech recorded in a variety of conditions with limited amounts of transcription." [1] The Open Keyword Search Evaluation was held in April 2013 (OpenKWS13) to the public using the surprise language of Vietnamese[2]. The keyword search system described in this work is based on our submission to OpenKWS13 with certain alterations, refinements, and elaborations.

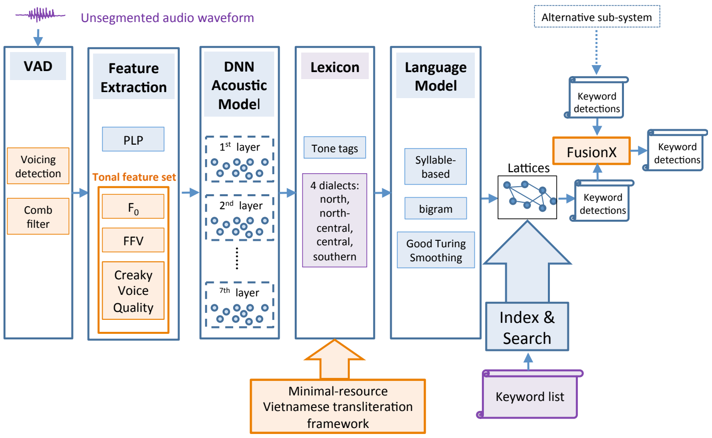

**Fig. 1**. Proposed keyword search system for conversational Vietnamese: orange blocks developed in-house (voice activity detection (VAD), tonal feature set, transliteration, FusionX); purple blocks were distributed to OpenKWS13 participants from NIST (lexicon, keyword list); blue blocks are from open-source (e.g., Kaldi [4]).

## 2. RELATION TO PRIOR WORK

Fig. 1 is an overview of the proposed KWS system. In our transcribe-and-search paradigm, a hybrid DNN-HMM based ASR system is first built using the Kaldi toolkit [6]; rule-based transliteration is used to introduce foreign out-of-vocabulary terms into the lexicon for decoding. Search takes place over the lattices generated by the ASR system. Our system differs from prior work in Vietnamese KWS in the following ways:

(1) Creaky voice quality features: To model lexical tones in Vietnamese, in addition to conventional tonal features such as pitch, we exploit features for creaky voice quality, which characterizes the *broken* tone peculiar to the northern Vietnamese dialect. Creaky voice quality features have been used to identify English allophones [5]. The glottalization phoneme (a short period of creaky voice) in Tagalog has been modeled by exploring different hidden Markov model topologies and using pitch/voicing features [6]. In Vietnamese ASR [7], only traditional pitch features have been used to model tones.

[1] IARPA broad agency announcement IARPA-BAA-11-02, 2011.

[2] Overview of the NIST Open Keyword Search 2013 Evaluation Workshop: http://www.signalprocessingsociety.org/technical-committees/list/sl-tc/spl-nl/2013-08/

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

To the best of our knowledge, no work to date has exploited creaky voice quality features in combination with pitch-related features.

(2) Minimal-resource transliteration framework: To alleviate the effect on out-of-vocabulary (OOV) keywords originating from non-Vietnamese (e.g., *facebook, android*), we propose a rule-based transliteration system. While OOVs have been studied a lot in KWS [8, 9], most work have not focused on OOVs of foreign origin. Statistical methods for handling OOVs (e.g., [10], [11], [12]) often suffer from performance drop when training data is limited, which is especially true for foreign-origin OOVs for low-resourced languages. To the best of our knowledge, while there has been limited work adopting standard statistical transliteration models for Vietnamese [13], to date no rule-based transliteration system has been proposed for Vietnamese elsewhere.

(3) System combination scheme FusionX: Many KWS systems employ system combination techniques to exploit diverse and complementary subsystems to improve performance (e.g., [14], [15]). Our proposed FusionX approach is in spirit similar to CombMNZ [16] used in [17]. The main difference is that our proposed approach does not need score normalization before combination.

## 3. HIGHLIGHTS OF PROPOSED KWS SYSTEM

In Fig. 1, speech is first segmented with a voice activity detector, and then features are extracted to train a SGMM-MMI [18] acoustic model. The target classes for training the DNN are obtained from forced-alignments of the training data using the SGMM-MMI model. Lexical entries contain tonal tags of four dialects (i.e., Northern, North-Central, Central, and Southern). A bigram language model is estimated from the transcriptions of the acoustic training data. A weighted finite state transducer based decoder is used to generate indexible lattices from the test data. The lattices are indexed and searched for specified keywords. Three subsystems were built and the keyword search results from each system is combined with FusionX to output a final result. We further describe our in-house components in the sections following.

### 3.1. Voice activity detection

A comb filter was used to detect peaks of voicing regions, which were then smoothed to fill in the gaps to determine the voice activity regions [19] of the evaluation dataset.

### 3.2. Tonal feature set

We used perceptual linear prediction (PLP) features [20] in our baseline system for the acoustic model of 67 phones. Since there are 6 distinct lexical tones in Vietnamese, 6 tone tags were used as questions during decision tree clustering. To better characterize these tones, we propose the following tonal feature set.

**(1) Fundamental Frequency Estimate $F_0$**
The fundamental frequency $F_0$ was estimated using a standard algorithm implemented using the ETSI standard[3], based on computing the magnitude of short time Fourier transform and interpolating the first spectral peak for voiced frames. Unvoiced frames were given a default value close to zero, and further post processing by a moving average filter smoothening the interpolated pitch values. This gives us a frame-by-frame estimate for F0, yielding a single scalar representation of pitch in each frame for unvoiced frames [21].

**(2) Fundamental Frequency Variation (FFV) features**
In contrast to standard scalar-valued representations of pitch, fundamental frequency variation (FFV) [22] [23] is a vector representation of delta pitch. It applies two asymmetric windows on the same frame, one window emphasizes earlier samples and the other emphasizes later samples, and effectively compares their spectra with different assumptions for the rate of pitch variation to characterize delta pitch: the different assumptions for delta pitch contribute to each element in the FFV spectrum.

The spectrum is further reduced by applying a 7-output filterbank with trapezoidal filters centered around different rates of frequency variation: one for perceptually flat pitch, two for slowly varying pitch, two for rapidly changing pitch. Finally, two rectangular filters at the extremities are used to handle unvoiced frames, which tend to have flat rather than decaying tails in the FFV spectrum. Thus, unlike F0 which does not define features for unvoiced regions, the FFV can characterize such regions as well, giving a 7-dimension vector for each speech frame.

**(3) Creaky Voice Quality (CVQ) features**
The *broken* tone (specified with the tilde diacritic in Vietnamese text, as in "$\tilde{a}$") in the northern Vietnamese dialect possesses *creaky* voice quality [24]. Creaky voice is caused by strong glottal constriction and is referred to as *strangled voice* (voix étranglée) in early French studies of the Vietnamese tone system [25]. This creakiness characteristic (sometimes referred to as glottalization) is peculiar to Vietnamese lexical tone productions.

Features such as pitch and FFV might not fully capture the acoustic characteristics of these creaky tonal productions. To this end, we adopt two features: (1) the 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 [26], and (2) the mean autocorrelation ratio [5], predicts the periodicity of glottal pulses.

H1$^*$-H2$^*$ has traditionally been viewed as one of the most important features in characterizing creaky voice. It is a spectral measure, which is potentially more prone to corruption of the telephone channel. On the other hand, the mean autocorrelation ratio (of the speech signal over a window containing at least 4 glottal pulses) is a temporal measure and less prone to the telephone channel [5]. In this work, CVQ features were extracted the same way as in [5], where regions with undefined values were padded with zero[4].

### 3.3. Minimal-resource transliteration system

#### *3.3.1. Proposed framework*

We developed a transliteration framework based on the training set to handle OOV terms of foreign origin. This algorithm is further refined with linguistic analysis in [27].

1. **Text-to-phoneme conversion:** We assume that the foreign loan word is in English and thus transliterate it to an English phonetic representation. This assumption is motivated by the fact that a large portion of loan words are from English, and that Vietnamese human transliterators often assume the foreign text follows English orthography and phonology because they are most familiar with English as their 2nd language.

[3] ETSI Standard: ETSI ES 202 212 V1.1.2, 2005.

[4] We thank Tae-Jin Yoon and Mark Hasegawa-Johnson for sharing their feature extraction script with us.

2. **Syllable splitting**

    (a) **Identify vowel nuclei**: The anchor points of each syllable are specified using vowels. If there are two consecutive vowels, the syllable boundary is set in between the vowels. Otherwise, the default is to structure syllable initials with consonants.

    (b) **Locate syllable-initial consonants**: All possible syllable-initial consonants are located to maximize the number of syllables with initial consonants (since 90.09% of native Vietnamese syllables start with a consonant in the training set).

    (c) **Locate syllable-final consonants**: Syllable-final consonants are located to complete the syllable if needed.

3. **Foreign phoneme to Vietnamese phoneme mapping**: The mappings were determined by the Language Specific Peculiarities Document of OpenKWS13 and [27].

4. **Tone specification** If the syllable ends with stop consonants /p, t, k/, Tone 2 (rising tone, specified with the *acute* diacritic in Vietnamese text, such as "á") is assigned, else Tone 1 (level tone, where no diacritic is marked as in "a") is assigned [27].

### *3.3.2. Comparison with statistical-approach baselines*

Statistical transliteration approaches typically require more training data to perform well. In the context of limited linguistic resources, we empirically show that our rule-based transliteration system performs better than classic statistical systems.

Foreign words in the training lexicon were extracted and split into the training and developmental set, while foreign words in the developmental data were designated as the test set.

Table 1 shows that the proposed system M3 consistently outperforms the baselines: it improves system M1 by 14.73% relative in token error rate (TER) and by 15.44% relative in string error rate (SER: any token error is a string error); it improves system M2 by 9.84% relative in TER and by 4.56% relative in SER.

**Table 1**. Transliteration performance comparison: Improved transliteration performance from rule-based approach. TER: token error rate; SER: string error rate.

| System | Model | TER (%) | SER (%) |
|---|---|---|---|
| M1 | IBM noisy channel model [12] | 25.8 | 69.3 |
| M2 | Joint source channel model [10] | 24.4 | 61.4 |
| M3 | Proposed model | 22.0 | 58.6 |

## 3.4. System combination: FusionX

FusionX shares a similar spirit to methods such as CombMNZ [16]. FusionX is performed iteratively, combining two systems at a time. Given a main system and an auxiliary system, the fused system is an enhanced version of the main system.

For keyword $kw$, let $T_{main}$ and $S_{main}$ be the begin time-point and score of detection $D_{main}$ from the main system. Similarly, $T_{aux}$ and $S_{aux}$ are the begin time-point and score of detection $D_{aux}$ from the auxiliary system. $D_{aux} \in$ *Yes*, where $S_{aux} > \theta_d$ are integrated into the main system via two modes:

1. Merging: $|T_{main} - T_{aux}| \leq T_{gap}$

    (a) Time boundaries of the merged decision are averaged across the two systems, replacing those of $D_{main}$.

    (b) If decisions $D_{main}$ and $D_{aux}$ are both $Yes$, the scores of the merged decision are averaged across the two systems, replacing those of $D_{main}$.

    (c) If decision $D_{main}$ is *No* but decision $D_{aux}$ is *Yes*, $S_{aux}$is included in the main system.

2. Insertion: $|T_{main} - T_{aux}| > T_{gap}$
   Time boundaries of $D_{aux}$ and $S_{aux}$ are inserted into the main system.

After iterating through every keyword, the final updated main system is the fused system. The values of the threshold $\theta_d$ and the time gap $T_{gap}$ are tuned on the developmental set. Default settings: $T_{gap} = 0.6$, $\theta_d = 0.75 - 0.9$.

# 4. EXPERIMENTS

## 4.1. OpenKWS13 Corpus

The training set includes 80 hours of conversational telephone speech with word transcriptions and 20 hours of scripted telephone speech. The developmental set is 10 hours; the evaluation set is 75 hours with no transcriptions nor timing information. The corpus covers various acoustic modeling challenges such as spontaneous speaking styles (e.g., hesitations), dialect diversity (i.e., Northern, Northern-central, Central, Southern), channel mismatch.

## 4.2. Automatic speech recognition (ASR) experiment

### *4.2.1. Implementation details*

*Acoustic model:* The baseline system S1 has 13-dimensions of PLP features, where 9 frames were concatenated together to apply LDA, MLLT, and fMLLR transforms, resulting in 40-dimensions. System S2 and S3 are similarly set up, differing only in the raw feature set and the LDA dimension. The DNN system has 7 hidden layers with 2048 nodes each, with almost similar number of senones. This information is summarized in Table 2. The training procedure is divided into three steps: (1) 10 iterations of pre-training; (2) training with cross entropy criterion; (3) scalable minimum Bays risk criterion based sequence training [28].

*Language model:* We used SRILM [29] to train a syllable-based bigram LM with Good Turing smoothing after exploring different configurations (e.g., 2-4gram; (multi-)syllable units).

### *4.2.2. Results: tonal features reduce word error rate*

Results are shown in Table 2 in word error rate (WER). The best system is S2, reaching WER=52%. Adding tonal features to PLP helps improve ASR performance: S2 ($F_0$ and FFV features) improves the baseline S1 (PLP features) by 6.47% relative (3.6% absolute); S3 ($F_0$, FFV, and CVQ features) improves the baseline S1 (PLP features) by 5.94% relative (3.3% absolute). Though S3 improves S1 slightly less than S2, the combination of S2 and S3 reaches fusion gains of 7.54% relative (see Section 4.3.3), implying that CVQ features complement pitch-related features for KWS tasks.

## 4.3. Keyword search (KWS) experiment

### *4.3.1. Implementation details*

*ASR Lattice indexing and search:* We used exact inverted indexes for lattices with timing information [30] to maintain search complexity to be linear to query length. Deterministic weighted finite state transducer were used to store soft-hits, containing the utterance id, start/end time, and posterior score.

**Table 2**. ASR acoustic model setup and individual system performance comparison. PLP: perceptual linear prediction; FFV: fundamental frequency variation; CVQ: creaky voice quality. Numbers in subscripts in column 2 specify the dimensions for each individual feature.

| System | Features | Feature dimension | LDA Dimension | No. of Senones | WER(%) | ATWV | MTWV |
|---|---|---|---|---|---|---|---|
| S1 | $PLP_{13}$ | 13 | 40 | 8,584 | 55.6 | 0.3892 | 0.3929 |
| S2 | $PLP_{13} + F_{0_1} + FFV_7$ | 21 | 63 | 8,642 | 52.0 | 0.4070 | 0.4148 |
| S3 | $PLP_{13} + F_{0_1} + FFV_7 + CVQ_2$ | 23 | 69 | 8,605 | 52.3 | 0.4051 | 0.4098 |

*Score normalization:* We adapt keyword-specific threshold normalization in [31] so the common threshold after normalization is 0.5.

*Evaluation metric:* Term-weighted value (TWV), which is 1 minus the weighted sum of the term-weighted probability of miss detection $P_{miss}(\theta)$ and the term-weighted probability of false alarm $P_{FA}(\theta)$:

$$TWV(\theta) = 1 - [P_{miss}(\theta) + \beta P_{FA}(\theta)], \qquad (1)$$

where $\theta$ is the decision threshold of the system. Actual term-weighted value (ATWV) is the TWV of the chosen decision threshold of the proposed system, whereas the maximum term-weighted value (MTWV) is the best TWV found over all $\theta$. While we report both measures, we only discuss and compare ATWV.

### 4.3.2. Foreign loan word keyword search task

We compiled a list of 140 keywords from the developmental set to specifically evaluate how well the proposed transliteration framework characterizes pronunciations adopted from foreign languages. The data partition is the same as in Section 3.3.2. System S2 was run using lexicons derived from four different transliteration models. As shown in Table 3, condition T1, T2, and T3 differ in where the pronunciations of the foreign keywords are derived from. Condition T3 (proposed model) outperforms condition T1 and T2 by 36.76% relative (2.11% absolute) and 14.93 % (1.02% absolute), respectively. Since the proposed transliteration model achieves the best performance, we adopt it in our KWS system.

**Table 3**. System S2 adopting various transliteration models.

| Condition | Transliteration model | ATWV | MTWV |
|---|---|---|---|
| T1 | IBM noisy channel model | 0.0574 | 0.0710 |
| T2 | Joint source channel model | 0.0683 | 0.0766 |
| T3 | Proposed model | 0.0785 | 0.0927 |

### 4.3.3. OpenKWS13 evaluation keyword search task

The OpenKWS13 evaluation keyword list contains 4,065 queries. Empirical results are shown in Table 2 and Table 4, where all ATWV results exceed the IARPA benchmark $ATWV = 0.3$. We also compare our system performance with the Babel teams in Table 5, where the WER of our system was obtained using segmental minimum Bayes risk decoding for lattice combination [32]. Below we summarize the trends of our results:

**(1) Tonal features complement PLP features.**
When comparing single system performance, we see that systems with tonal features (S2 and S3) outperform the baseline S1 system: S2 outperforms S1 by 4.6% relative (1.8% absolute); S3 outperforms S1 by 4.1% relative (1.6% absolute). When comparing system S1 with the fusion of S1+S2+S3, we see a relative gain of 21.83% (8.5% absolute). These results empirically validate the importance of tonal features in Vietnamese keyword search.

**Table 4**. KWS system combination results using FusionX.

| System | ATWV | MTWV |
|---|---|---|
| S1+S2 | 0.4447 | 0.4448 |
| S2+S3 | 0.4377 | 0.4398 |
| S1+S3 | 0.4581 | 0.4585 |
| S1+S2+S3 | 0.4742 | 0.4746 |

**Table 5**. Comparing proposed system performance (SINGA team) with Babel teams. All results are up to date.

| | Team name (lead institution) | WER | ATWV |
|---|---|---|---|
| Babel | BABELON (BBN) | 45.0 | 0.625 |
| | LORELEI (IBM) | 52.1 | 0.545 |
| | RADICAL (CMU) | 51.0 | 0.452 |
| | SWORDFISH (ICSI) | 55.9 | 0.461 |
| | SINGA ($I^2R$) | 50.0 | 0.474 |

**(2) Creaky voice quality features complement PLP features.**
S1+S2 outperforms S1 by 14.3% relative (5.6% absolute); S1+S3 outperforms S1 by 17.7% relative (6.9% absolute). Though S3 does not perform better than S2 as a single system, the fusion gains S3 brings to S1 are consistently larger than what S2 brings to S1. These empirical results show that creaky voice quality features are helpful in Vietnamese keyword search.

**(3) Creaky voice quality features complement $F_0$ and FFV.**
We see that although S2 and S3 perform similarly as single systems in Table 2, fusion gain reaches 7.54% relative (3.07 % absolute), when they are combined (S2+S3 in Table 4). *This gain implies that creaky voice quality features model tonal attributes that differ from absolute pitch estimates and pitch variation estimates.*

**(4) Fusing all 3 sub-systems achieves best performance.**
The best fused system (S1+S2+S3) outperforms the best single system S2 by 16.51% relative (6.72% absolute), and the baseline system S1 by 21.83% (8.5% absolute).

## 5. DISCUSSION

We presented strategies for a state-of-the-art Vietnamese keyword search (KWS) system. Highlights of the KWS system include: (1) acoustic features characterizing creaky voice quality peculiar to lexical tones in Vietnamese, (2) a minimal-resource transliteration framework, and (3) a proposed system combination scheme FusionX. The ATWV of the proposed system reaches at least 0.47, exceeding the ATWV=0.3 benchmark for IARPA Babel participants in the NIST OpenKWS13 Evaluation.

Our proposed KWS system was built on open-source tools (e.g., Kaldi [4]) augmented with enhancements developed in-house.Thus researchers new to KWS can readily set-up competitive baseline systems by adopting our work. For future work, we plan to apply and adapt the proposed strategies to other under-resourced languages.

## 6. REFERENCES

[1] Joseph Keshet, David Grangier, and Samy Bengio, “Discriminative keyword spotting,” *Speech Communication*, vol. 51, no. 4, pp. 317–329, 2009.

[2] I-Fan Chen and Chin-Hui Lee, “A hybrid HMM/DNN Approach to keyword spotting for short words,” in *Proc. of Interspeech*, 2013, pp. 1574 – 1578.

[3] Jonathan G Fiscus, Jerome Ajot, John S Garofolo, and George Doddingtion, “Results of the 2006 spoken term detection evaluation,” in *Proceedings of ACM SIGIR Workshop on Searching Spontaneous Conversational*, 2007, pp. 51–55.

[4] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., “The kaldi speech recognition toolkit,” in *Proc. of IEEE ASRU*, 2011.

[5] Tae-Jin Yoon, Xiaodan Zhuang, Jennifer Cole, and Mark Hasegawa-Johnson, “Voice quality dependent speech recognition,” in *International Symposium on Linguistic Patterns in Spontaneous Speech*, 2008.

[6] Korbinian Riedhammer, Van Hai Do, and James Hieronymus, Eds., *A Study on LVCSR and Keyword Search for Tagalog*. Proc. of Interspeech, 2013.

[7] Ngoc Thang Vu and Tanja Schultz, “Vietnamese large vocabulary continuous speech recognition,” in *Proc. of IEEE ASRU*. IEEE, 2009, pp. 333–338.

[8] Murat Akbacak, Dimitra Vergyri, and Andreas Stolcke, “Open-vocabulary spoken term detection using graphone-based hybrid recognition systems,” in *Proc. of IEEE ICASSP*, 2008, pp. 5240–5243.

[9] Dogan Can, Erica Cooper, Abhinav Sethy, Chris White, Bhuvana Ramabhadran, and Murat Saraclar, “Effect of pronounciations on OOV queries in spoken term detection,” in *Proc. of IEEE ICASSP*, 2009, pp. 3957–3960.

[10] Li Haizhou, Zhang Min, and Su Jian, “A joint source-channel model for machine transliteration,” in *Proc. of the 42nd Annual Meeting on Association for Computational Linguistics*, 2004.

[11] Maximilian Bisani and Hermann Ney, “Joint-sequence models for grapheme-to-phoneme conversion,” *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008.

[12] Warren Weaver, “Translation,” *Machine translation of languages*, vol. 14, pp. 15–23, 1955.

[13] Nam X Cao, Nhut M Pham, and Quan H Vu, “Comparative analysis of transliteration techniques based on statistical machine translation and joint-sequence model,” in *Proc. of Symposium on Information and Communication Technology*. ACM, 2010, pp. 59–63.

[14] Lidia Mangu, Hagen Soltau, Hong-Kwang Kuo, Brian Kingsbury, and George Saon, “Exploiting diversity for spoken term detection,” in *Proc. of IEEE ICASSP*, 2013.

[15] Alberto Abad, Luis Javier Rodrıguez-Fuentes, Mikel Penagarikano, Amparo Varona, and Germán Bordel, “On the calibration and fusion of heterogeneous spoken term detectionsystems,” in *Proc. of Interspeech*, 2013.

[16] E. A. Fox and J. A. Shaw, “Combination of multiple searches,” *NIST Special Publication SP*, pp. 243–243, 1994.

[17] Brian Kingsbury, Jia Cui, Xiaodong Cui, Mark JF Gales, Kate Knill, Jonathan Mamou, Lidia Mangu, David Nolden, Michael Picheny, Bhuvana Ramabhadran, Ralf Schluter, Abhinav Sethy, and Phillip C. Woodland, “A High Performance Cantonese Keyword Search System,” in *Proc. ICASSP*, 2013.

[18] Daniel Povey, Lukas Burget, Mohit Agarwal, Pinar Akyazi, Kai Feng, Arnab Ghoshal, Ondrej Glembek, Nagendra K Goel, Martin Karafiát, Ariya Rastrow, R. C. Rose, P Schearz, and S. Thomas, “Subspace gaussian mixture models for speech recognition,” in *IEEE International Conference on Acoustics Speech and Signal Processing,*. IEEE, 2010, pp. 4330–4333.

[19] Nguyen Trung Hieu Shengkui Zhao, Eng Siong Chng and Haizhou Li, “A Robust Real-time Sound Source Localization System for Olivia Robot,” in *2010 APSIPA Annual Summit and Conference*, 2010.

[20] Hynek Hermansky, “Perceptual linear predictive (PLP) analysis of speech,” *J. Acoust. Soc. Am.*, vol. 87, pp. 1738, 1990.

[21] C Julian Chen, Ramesh A Gopinath, Michael D Monkowski, Michael A Picheny, and Katherine Shen, “New methods in continuous mandarin speech recognition.,” in *Proc. of Eurospeech*, 1997.

[22] Kornel Laskowski and Qin Jin, “Modeling instantaneous intonation for speaker identification using the fundamental frequency variation spectrum,” in *Proc. of IEEE ICASSP*, 2009, pp. 4541–4544.

[23] Florian Metze, Zaid A. W. Sheikh, Alex Waibel, Jonas Gehring, Kevin Kilgour, Quoc Bao Nguyen, and Van Huy Nguyen, “Models of tone for tonal and non-tonal languages,” in *Proc. IEEE ASRU*, Olomouc; Czech Republic, 2013.

[24] Do Dat Tran, Eric Castelli, Jean-François Serignat, Van Loan Trinh, and Le Xuan Hung, “Influence of F0 on Vietnamese syllable perception,” in *Proc. of Interspeech*, 2005, pp. 1697–1700.

[25] Vu Ngoc Tuan, Christophe d’Alessandro, and Sophie Rosset, “A phonetic study of Vietnamese tones: acoustic and electroglottographic measurements,” in *Proc. of Interspeech*, 2002.

[26] Helen M Hanson, “Glottal characteristics of female speakers: Acoustic correlates,” *J. Acoust. Soc. Am.*, vol. 101, pp. 466, 1997.

[27] Thi-Quynh-Hoa Hoang, “A phonological contrastive study of Vietnamese and English,” MA Thesis, Texas Technology College, 1965.

[28] L. Burget K. Vesely, A. Ghoshal and D. Povey, “Sequence-discriminative training of deep neural networks,” in *Proceedings of Interspeech*, 2013.

[29] Andreas Stolcke et al., “SRILM-an extensible language modeling toolkit,” in *Proc. of Interspeech*, 2002.

[30] Dogan Can and Murat Saraclar, “Lattice indexing for spoken term detection,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2338–2347, 2011.

[31] David RH Miller, Michael Kleber, Chia-Lin Kao, Owen Kimball, Thomas Colthurst, Stephen A Lowe, Richard M Schwartz, and Herbert Gish, “Rapid and accurate spoken term detection.,” in *Proc. of Interspeech*, 2007, pp. 314–317.

[32] Vaibhava Goel, Shankar Kumar, and William Byrne, “Segmental minimum Bayes-risk decoding for automatic speech recognition,” *Speech and Audio Processing, IEEE Transactions on*, vol. 12, no. 3, pp. 234–249, 2004.