SHEFCE: A CANTONESE-ENGLISH BILINGUAL SPEECH CORPUS FOR PRONUNCIATION ASSESSMENT

Raymond W. M. Ng[1]*, Alvin C.M. Kwan*[2]*, Tan Lee*[3] *and Thomas Hain*[1]

[1] Department of Computer Science, University of Sheffield, United Kingdom,
[2] Faculty of Education, The University of Hong Kong, Hong Kong,
[3] Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong.

ABSTRACT

This paper introduces the development of ShefCE: a Cantonese-English bilingual speech corpus from L2 English speakers in Hong Kong. Bilingual parallel recording materials were chosen from TED online lectures. Script selection were carried out according to bilingual consistency (evaluated using a machine translation system) and the distribution balance of phonemes. 31 undergraduate to postgraduate students in Hong Kong aged 20-30 were recruited and recorded a 25-hour speech corpus (12 hours in Cantonese and 13 hours in English). Baseline phoneme/syllable recognition systems were trained on background data with and without the ShefCE training data. The final syllable error rate (SER) for Cantonese is 17.3% and final phoneme error rate (PER) for English is 34.5%. The automatic speech recognition performance on English showed a significant mismatch when applying L1 models on L2 data, suggesting the need for explicit accent adaptation. ShefCE and the corresponding baseline models will be made openly available for academic research.

***Index Terms*—** Bilingual parallel speech corpus, Cantonese, English pronunciation assessment

1. INTRODUCTION

Second language acquisition is a complex cognitive process that involves the learning of pronunciation, grammar, semantics and usage. To focus on the pronunciation aspect, when speaking a second language (L2), a speaker's pronunciation is often different from that of a native speaker because the phonetic space and the co-occurrence statistics of phonemes in two languages are often very different, and it is difficult for most speakers to adapt this difference when speaking in L2 [1]. Some of these pronunciation discrepancies give rise to pronunciation errors, while others which do not affect communication can be regarded as accents.

In terms of language learning, it is important to locate pronunciation errors and tolerate accents. Nevertheless, even for professional teachers the distinction between accents and pronunciation errors is not trivial, and in many cases it is down to a subjective judgement [2]. Automatic methods can be useful in giving objective judgement and suggestions [3]. With spoken language technology, statistic-based methods and objective measurements may provide a useful pedagogical reference for language learning [4, 5]. Phone-level pronunciation is an important assessment index and it can be evaluated with a variety of metrics [6]. Annotations on pronunciation quality and non-native L2 accented data were used to train models for pronunciation quality prediction [7].

There were several studies relevant to English learners in Hong Kong whose mother tongue is Cantonese. In [4], a multimedia computer-aided language learning tool was designed for this target learner group. The ICNALE-Spoken corpus contains L2 English spoken by Asian nationals including Hong Kong speakers [8]. Given the huge difference between the Cantonese and English phonology, a study on the phonetic transfer between L1 and L2 would help to understand language learning mechanism. In [9], characteristics of L1 (mother tongue) was considered to enhance the correlation between expert and automatic pronunciation scoring.

This research focuses on the pronunciation of L2 English by Hong Kong English learners. The availability of data is critical to pronunciation assessment. As a first step to address this problem, we constructed ShefCE — a parallel bilingual Cantonese-English speech corpus from the English learners in Hong Kong with a total duration of 25 hours (13 hours in English, 12 hours in Cantonese). Baseline automatic speech recognisers were trained to quantify the capability of machine processing. As for related studies, similar effort on bilingual corpus construction was seen in the French-German language pair [10]; there was also a multi-dialect Arabic parallel speech corpus found in the literature [11]. To our knowledge, no such dataset is available for Cantonese and English.

This paper presents ShefCE — a Cantonese English bilingual parallel speech corpus which can be used for pronunciation assessment studies. In this paper, the data collection process is explained. Baseline experiments of automatic syllable and phoneme recognition on Cantonese and English data are presented to demonstrate the quality of ShefCE data. The English phoneme recognition experiment results will shed light on the language mismatch issues when automatic speech recognition is applied on L2 speech.

2. BILINGUAL PARALLEL DATA SELECTION

ShefCE aims at creating a parallel bilingual speech dataset in Cantonese and English primarily for language learning. By having a parallel corpus, interlanguage transfer on the articulatory level could be studied for pronunciation assessment. Potentially, analysis of different linguistic aspects in language learning could also be done on the grammatical and semantic levels. The target recording materials were chosen from the TED website [12], which hosts freely viewable public lectures in various topics such as science, culture, humanity, etc. Most of the lectures were delivered in English, with subtitles available. Some lectures were also translated into Traditional Chinese, the writing script for (Hong Kong) Cantonese. The transcripts

This work is partially funded by Impact, Innovation and Knowledge Exchange (IIKE) Fund, University of Sheffield and Google. Special thanks to Helen Meng, CUHK, for her special support and advice on corpus collection and project preparation.

were segmented at the utterance level.

To source the recording material, the transcripts of TED talks with talk ID smaller than 1000 were downloaded from the TED website. Preliminary filtering removed the talks where parallel bilingual transcripts were not available. This left 683 talks, with each of them having 290 utterances on average.

To ensure a certain quality standard of the crawled recording materials, cross-lingual relationship of the parallel text and phonetic balance of the English text were analysed. To study the cross-lingual relationship, a separate set of 898 talks (with talk ID larger than 1000) was downloaded from TED. A Chinese-to-English phrase-based translation system was built using the open-sourced toolkit MOSES [13]. In the implementation, we followed the details of an analogous English-French system described in [14]. For Chinese word segmentation, the Stanford Chinese segmenter, an off-the-shelf Chinese word segmentation tool, was used [15]. The Chinese transcripts of 683 talks were automatically translated to English using the trained model and translation performance per talk was evaluated in terms of METEOR scores. Higher METEOR scores, indicating better translation quality, were preferred as they represented higher consistencies between the bilingual parallel transcripts.

For English phonetic balance, the English transcripts were converted to phoneme labels using a dictionary containing 60k words. Subsequently, word and utterance boundaries were disregarded and context-dependent phoneme statistics for each talk was computed and represented in fixed length term frequency-inverse document frequency (tf-idf) vectors. Take triphone as an example, let tf-idf(i) denote the relevant statistics of the i^{th} triphone in the vocabulary and the total number of triphones is I. A cross-entropy metric CE is computed to represent the phonetic variety,

$$CE = -\sum_{i}^{I} \frac{1}{I} log(\text{tf-idf}_{\text{norm}}(i)), \quad (1)$$

where $\text{tf-idf}_{\text{norm}}(i) = \frac{\text{tf-idf}(i)}{\sum_i \text{tf-idf}(i)}$. This computation was repeated with triphone, biphone and monophone statistics. A high CE value indicates a more balanced distribution of the target units (triphones / biphones or monophones) and thus is more desirable.

We selected the talks to be recorded according to the METEOR score, monophone, biphone and triphone cross entropy statistics. Out of the 683 talks, those with a METEOR score above 15 were selected. This resulted in 512 talks. After this, a manual inspection of all four statistics was conducted among the selected talks. Talks with very high METEOR scores and inconsistent CE among monophones, biphones and triphones were removed. A closer inspection indicated that removed talks contained repeated scripts (lyrics or songs), or repeated short annotations (e.g. [MUSIC], [APPLAUSE] instead of meaningful subtitles). Finally, a total of 40 talks was chosen for bilingual speech recording.

Table 1. List of talks in Train and Test set. f denotes full talk, fraction denotes partial talk (e.g. $\frac{2}{3}$ denotes 2$^{\text{nd}}$ part in a 3-part talk

Set	Speakers	Talks
Training	all except 0002(F), 0009(F), 0014(F), 0021(M),0022(M), 0028(M)	2($\frac{1}{3}$), 14(f), 42($\frac{1}{3}$,$\frac{2}{3}$), 46(f), 72($\frac{1}{2}$,$\frac{2}{2}$), 79($\frac{1}{3}$),112($\frac{1}{5}$), 117($\frac{3}{3}$), 126($\frac{2}{2}$), 128($\frac{1}{3}$,$\frac{2}{3}$), 139(f), 165(f), 183(f)$^{\#}$, 197($\frac{1}{2}$), 201(f), 207($\frac{1}{3}$,$\frac{2}{3}$), 211(f,$\frac{1}{2}$,$\frac{2}{2}$), 224(f), 225(f), 248($\frac{1}{2}$), 252(f), 306($\frac{1}{3}$,$\frac{2}{3}$,$\frac{3}{3}$), 313(f), 407(f), 411($\frac{1}{2}$), 416($\frac{1}{3}$,$\frac{2}{3}$), 464($\frac{1}{2}$,$\frac{2}{2}$), 489(f), 492(f), 657(f), 662(f), 680($\frac{1}{2}$,$\frac{2}{2}$), 684(f), 965(f,$\frac{1}{3}$,$\frac{3}{3}$)
Test	0002(F), 0009(F), 0014(F), 0021(M), 0022(M), 0028(M)	2($\frac{2}{3}$), 42($\frac{3}{3}$), 47($\frac{1}{2}$,$\frac{2}{2}$), 79($\frac{2}{3}$), 123(f), 126($\frac{1}{2}$), 177($\frac{2}{3}$), 183(f)$^{\#}$, 351($\frac{1}{2}$), 501(f), 901(f).

$^{\#}$: The only overlapping talk (id : 183) between Training and Test sets

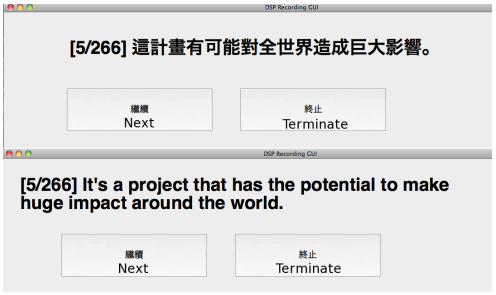

Fig. 1. Screen capture of the interactive recording programme for Cantonese (above) and English (below)

To avoid subject turning fatigue in prolonged recording sessions and to attain a better control of recording time, long talks were truncated (segmented in 2 to 5 parts). Training and test sets were designed in such a way that truncated talks and speakers do not overlap. The only exception is Talk 183 which appeared in both sets. This acts as a control for future experiments. Table 1 shows the list of truncated talks in ShefCE. There are a total of 47 and 12 truncated talks in the training and test set respectively.

3. BILINGUAL DATA RECORDING

20 (ID:0001-0020) subjects were recruited from the Faculty of Education in the University of Hong Kong (HKU) and 11 (ID:0021-0031) from the Department of Electronic Engineering in the Chinese University of Hong Kong (CUHK). The subjects are undergraduate to postgraduate students aged 20-30. Altogether there are 31 speakers, female to male ratio is 18 to 13. 3 male and 3 female speakers were chosen as test speakers. Table 2 shows the breakdown of speakers by gender, institutions and training-test set distribution.

All 31 subjects have Cantonese as their native language. 4 of the subjects grew up in China or overseas, despite Cantonese still being their mother tongue. 27 subjects received primary and secondary education in Hong Kong. From informal self report on language proficiency and subjective comments from the academic tutors of the subjects, it is known that the English language proficiency level of the 31 speakers vary. The social background and disparity of language proficiency essentially reflect the English language capabilities of Hong Kong students nowadays. As we did not intend to carry out fine-level linguistic analysis to individual speakers, subject to ethical requirements these potentially identifiable information was not retained with the corpus. We pooled all speakers in model training and the same quantitative methods were applied to all speakers.

To conduct the recording, the target recording scripts described in §2 were pre-processed and displayed on a computer screen in an utterance-by-utterance basis. An example screen prompt is shown in Figure 1. The subject were asked to read the prompt in Cantonese, or English, as displayed. After an utterance was read, the subject would click the "Next" button to proceed to the subsequent utterance. Timings of button clicks were tracked and used for utterance segmentation.

Recordings were made in two distinct environment and microphone settings. The HKU speakers (subject ID 0001-0020) were recorded in a quiet office space without professional soundproof fa-

Table 2. Gender, institution and data distribution in ShefCE corpus

	Training set		Test set	
Institution	Female	Male	Female	Male
HKU (ID:0001-0020)	15	2	3	0
CUHK (ID:0021-0031)	0	8	0	3

Table 3. List of Cantonese initials and finals

Type	List of subsyllables within the type
Initial	[null],b,c,d,f,g,gw,h,j,k,kw,l,m,n,ng,p,s,t,w,z
Final	aa,aai,aak,aam,aan,aang,aap,aat,aau,ai,ak,am,an, ang,ap,at,au,e,ei,ek,eng,eoi,eon,eot,i,ik,im,in, ing,ip,it,iu,m,ng,o,oe,oek,oeng,oi,ok,on,ong,ot, ou,u,ui,uk,un,ung,ut,yu,yun,yut

cilities. A laptop computer with 2.53GHz Intel Core 2 Duo CPU and 4GB memory were used for recording. On-board microphones were used and the noise cancellation algorithm of the operating system was applied. The CUHK speakers (subject ID 0021-0031) were recorded in a soundproof room with a Sputnik vacuum tube large-diaphragm condenser microphone with Cardioid polar pattern directing to the speakers. All recording were made in RIFF wav format with 16-bit PCM coding and 44.1kHz sampling.

To match the background data and model, all data were down-sampled to 16kHz before ASR models were trained and/or tested. The accuracy of reference transcripts in ShefCE data was verified manually. 3% of the training data (180 Cantonese and 175 English utterances) were selected. The mismatch between the reference transcripts and the actually spoken words/syllables was found out by manual listening. In terms of syllable and word error rates, the mismatch rate in Cantonese and English ShefCE data is 5.4% and 3.4% respectively.

4. PHONETIC UNITS IN ENGLISH AND CANTONESE

The primary purpose of the ShefCE corpus is for pronunciation assessment. For English data, model building focuses on the phone units. We used the common pronunciation dictionary with pronunciation variants included and stress levels omitted. There are a total of 39 phonemes.

Cantonese is a syllabic language. Each word is made up of one or more Chinese characters. 1-character and 2-character word accounts for the majority. There are roughly 2500 commonly used Chinese characters. Each Chinese character has 1 (or in minority homophonic cases, 2-3) monosyllabic pronunciation in Cantonese. Cantonese syllable follows a (C)V(C) structure. In speech recognition, acoustic modelling is done on the sub-syllable level. The syllable is segmented into two parts – "initial" and "final". "Initial" represents the first constant in the syllable (The optional absence of consonant prefix is denoted as a [null] initial). "Final" represents the middle vowel, which may correspond to long/short monophthongs or diphthongs. A syllable coda in the final is optional. There are in total 20 initial (including [null]) and 53 finals in Cantonese. Under sub-syllable unit combinatorial constraints the total number of Cantonese syllables is 689. Table 3 enumerates all Cantonese initial and final units. These are the basic phonetic units the Cantonese acoustic models learn [1].

Lexical tones in Cantonese are suprasegmental features which in most cases appear as redundant with the presence of contextual segmental features. Unlike other Cantonese speech recognition system setups where lexical tones are explicitly modelled [16, 17], we made use of linguistic knowledge and sub-syllable unit combinatorial constraints, constructed a refined lexicon and did not model lexical tones explicitly.

[1]Post-experimental studies showed that using a Cantonese phoneme model could improve syllable error rates by 0.2%-0.4% absolute.

5. SPEECH RECOGNITION SYSTEMS

Along with the data, baseline speech recognition models in Cantonese and English were trained and made available for further studies with this data set. Cantonese and English read speech data were sourced from existing data from the CUSENT [18] and WSJ0 [19] corpora to train *background models*. For each language, mixed-condition training was also carried out by mixing the background data with the ShefCE training data, to provide *mixed-condition models*.

Background and *mixed-condition* models differ only in the training data. Cantonese and English *background* models were trained on CUSENT (68 speakers, 19.4 hours) and WSJ0:SI-84 (83 speakers, 15.2 hours) data respectively. ShefCE training data were added to train *mixed-condition* models. The ShefCE training data contains 25 speakers. The total duration of Cantonese and English training data is 9.7 and 10.4 hours respectively.

Training of both background and mixed-condition models followed the same approach. Triphone GMM-HMM models were trained in maximum likelihood criterion followed by discriminative MPE training. Speaker-dependent feature MLLR (fMLLR) was learnt to transform the 13-dimensional MFCC plus 7-frame splicing (91-dimensional features) to 40 dimensions. Transform was learnt on the tied-state alignment (for training data) or decoding (for test data) using the MPE-trained GMM-HMM model. The fMLLR features further underwent ± 5 frame splicing to create 440-dimensional input feature to the DNN-HMM models. All DNN-HMM models have six layers, each having 2048 neurons. The network first went through Restricted Boltzmann Machine (RBM) unsupervised pretraining and then the network weights were fine tuned with the cross-entropy criterion. In the following, results using the MPE-trained GMM-HMM models, the speaker-adapted setting with fMLLR in GMM-HMM, and the DNN-HMM results will be reported.

No word-level language models were trained. For Cantonese data, a dictionary maps initial-final units to 689 syllables, a syllable bigram language model built on the training data was applied. For English data, a phoneme bigram language model was trained. The outputs of Cantonese and English recognition are syllables and phonemes respectively.

Cantonese and English speech recognition system were evaluated on syllable error rates (SER) and phoneme error rates (PER) respectively. The identity of Cantonese reference syllables can be determined in a rule-based method using a character-to-syllable lexicon. For English, reference phonemes cannot be determined in a trivial way. Upon every round of model training, automatic alignment using the latest model is used to decide the identity and time boundary of the phonemes, which is then compared with the decoding hypotheses to generate scoring results.

6. RESULTS

6.1. Background data

The background models were tested on CUSENT and WSJ0 background data (CUSENT:test, WSJ0:test-eval92) to benchmark the performance. As for CUSENT, the DNN-HMM model gave a syllable error rate (SER) of 7.52%.

For WSJ0 (test-eval92), experiment benchmark is available in the official Kaldi recipe with a SAT (speaker adaptive trained) GMM-HMM model trained only on SI-84 data [20]. With word trigram pruned language model, the official WER is 9.30%. This

is compared to our reported number of 8.95%. There are two differences between the Kaldi standard model and ours. First, we included additional discriminative training (minimum pronunciation error, MPE) before SAT. Second, the stress labels of phonemes in the lexicon were omitted.

Subsequently, the word trigram pruned language model was replaced by a phoneme bigram language model and the test data was decoded again. At the SAT stage, PER with and without stress-level modelling was 18.37% and 15.74%. The DNN-HMM models gave a PER 13.82% and 11.20% respectively with and without stress-level modelling. In the following, the stress labels of English phonemes would not be modelled and the phoneme bigram language model would be applied on English data.

6.2. Mixed-condition data

In Table 4, the Cantonese SER and English PER on ShefCE test data with the background (WSJ0, CUSENT) and the mixed-condition (WSJ0+ShefCE, CUSENT+ShefCE) models are reported. These include the results at three training stages:-

1. The GMM-HMM(MPE) models are discriminatively trained on minimum pronunciation error criterion.
2. The Speaker Adaptive Training (SAT) setting is based on feature-space MLLR (fMLLR). Initial training alignments were obtained from 1.
3. The DNN-HMM is the hybrid feed-forward neural network model. Initial training alignments were obtained from 2.

Across all training conditions, error rates decreased from GMM-HMM(MPE), fMLLR to DNN-HMM. The relative improvement from GMM-HMM(MPE) to DNN-HMM is consistent across background and mixed-condition models in Cantonese (28-29%). For English, an improvement of 17.1% was observed when using the background model, compared with 22.4% when the mixed-condition model was used. The latter is regarded to have captured the properties of accented data (uttered by L2 English learners).

Focusing on the DNN-HMM models, the Cantonese background model (CUSENT) gave an SER of 24.9% on ShefCE test data. On English, the DNN-HMM background model (WSJ0) gave a PER of 51.51% on ShefCE test data. Background models gave high error rates when they were applied on ShefCE English test data (§6.1 – as opposed to 11.2% when the same model was applied on background data). This is consistent with previous studies, where high error rates were observed due to the language mismatch between training on L1 and testing on L2 English speakers [4].

Mixed-condition DNN-HMM models gave 30% relatively lower error rates compared with using only the background models.

Table 5 shows the most frequent phoneme substitution patterns on ShefCE English test data with the DNN-HMM background model (trained on L1 data) and the DNN-HMM mixed-condition model (trained on L1 and L2 data). Common confusions occur between long and short vowels, and between voiced and unvoiced constants. In the mixed-condition model, the absolute occurrence of most confusion patterns decrease, leading to a lower phoneme error rate. Some of the pronunciation confusions might have reflected a consistent shift of pronunciation across all L2 speakers. For instance, pronouncing voiced labiodental fricatives (v) and voiced dental fricatives (th) as unvoiced labiodental fricative (f) is quite common for English L2 speakers in Hong Kong. In mixed-condition training the model has a clear direction to adapt this, leading to a significant drop of recognition error (th/v→f substitutions reduced from 338 to 171 times). A minority of confusion patterns is bilateral and

Table 4. Syllable error rate (SER) and phoneme error rate (PER) of background and mixed-condition training on ShefCE test data

	Cantonese SER with training data:		English PER with training data:	
	CUSENT	+ShefCE	WSJ0	+ShefCE
1. GMM-HMM(MPE)	35.0%	23.9%	62.1%	44.6%
2. SAT(fMLLR)	26.8%	21.6%	55.0%	41.2%
3. DNN-HMM	24.9%	17.3%	51.5%	34.5%
Relative improvement from 1. to 3	28.9%	27.6%	17.1%	22.4%

Table 5. Phoneme substitution on ShefCE(English) test data

Model (DNN-HMM)	Substitution patterns (occurrence > 0.2%)
Background model	s→z, z→s, ih→iy, d→t, dh→d, ae→eh, eh→ae, v→f, ah→ih, er→ah, ih→ah, dh→t, ah→ao, r→l, l→n, r→ah, l→ow, th→f, t→z, t→s, t→d, ah→ow, iy→ih
Mixed-condition model	s→z, z→s, ih→iy, eh→ae, d→t, er→ah, ih→ah, ae→ah, t→z, iy→ih, r→w, ah→ih, n→m, v→f, dh→d, d→dh, t→d

complicated. For example, s→z and z→s are highly confusable. Substitutions in two directions are 362 and 352 with the background model. With mixed-condition model, substitutions biased to s→z with 396 times (and 275 times for z→s). Mixed-condition acoustic models might have lower capabilities in recognising these phonemes. However, it may not have caused communication errors since phonotactic constraints in words would stand out and the phoneme bigram language model alone would be sufficient for phoneme classification. Further studies on the confusion patterns of different phonemes spoken by L2 speakers are believed to bring deeper understanding in accents and pronunciation errors by L2 speakers.

7. SUMMARY

In this paper, we discussed the preparation of a bilingual parallel speech corpus in Cantonese and accented English. Baseline models on Cantonese and English were trained and the results were presented in the paper. Mixed-condition training brought significant error rate reduction. Some specific phoneme substitution patterns were described. Further studies are necessary to bring deeper understanding on accents and pronunciation errors by L2 speakers. ShefCE is a speech corpus where the recording scripts are bilingual parallel text. This data is expected to help not only in the study of pronunciation assessment, but also in other areas such as multi-lingual speech recognition, spoken language translation, etc.

8. DATA ACCESS STATEMENT

Data used in this paper was obtained from these resources: CSR-I (WSJ0) Complete (LDC Catalogue No: LDC93S6A, ISBN: 1-58563-006-3, ISLRN: 296-840-353-630-9), CUSENT (`http://dsp.ee.cuhk.edu.hk/license_cucorpora.php`), TED (English transcript and Chinese translation, `http://www.ted.com`). The ShefCE corpus and the baseline Cantonese and English acoustic models are openly available for academic research and can be accessed online with DOI:`10.15131/shef.data.4522907` (corpus) and `10.15131/shef.data.4522925` (models).

9. REFERENCES

[1] Center for langauge education, The University of Science and Technology, Hong Kong, "Common pronunciation problems for cantonese speakers," http://cle.ust.hk/online_resources/common/advice/english/Spring%202011-12/Pronunciation/P7R.pdf, 2011.

[2] Tracey M. Derwing and Murray J. Munro, "Second language accent and pronunciation teaching: A research-based approach," *TESOL Quarterly*, vol. 39, no. 3, pp. 379–397, 2005.

[3] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832–844, Oct 2009.

[4] Brian Mak, Manhung Siu, Mimi Ng, Yik-Cheung Tam, Yu-Chung Chan, Kin-Wah Chan, Ka-Yee Leung, Simon Ho, Fong-Ho Chong, Jimmy Wong, and Jacqueline Lo, "Plaser: Pronunciation learning via automatic speech recognition," in *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing - Volume 2*, 2003, HLT-NAACL-EDUC '03, pp. 23–29.

[5] Mauro Nicolao, Amy V. Beeston, and Thomas Hain, "Automatic assessment of english learner pronunciation using discriminative training," in *Proc. ICASSP*, 2015.

[6] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2-3, pp. 95–108, Feb 2000.

[7] Tobias Cincarek, Rainer Gruhn, Christian Hacker, Elmar Nöth, and Satoshi Nakamura, "Automatic pronunciation scoring of words and sentences independent from the non-native's first language," *Computer Speech & Language*, vol. 23, no. 1, pp. 65 – 88, 2009.

[8] Shin'ichiro Ishikawa, "Design of the ICNALE-Spoken : A new database for multi-modal contrastive interlanguage analysis," *Learner Corpus Studies in Asia and the World*, vol. 2, pp. 63–75, 2014.

[9] N. Moustroufas and V. Digalakis, "Automatic pronunciation evaluation of foreign speakers using unknown text," *Computer Speech & Language*, vol. 21, no. 1, pp. 219 – 230, 2007.

[10] Camille Fauth, Anne Bonneau, Frank Zimmerer, Jurgen Trouvain Bistra Andreeva, Vincent Colotte, Dominque Fohr, Denis Jouvet, Jeanin Jügler, Yves Laprie, Odile Mella, and Bernd Möbius, "Designing a bilingual speech corpus for french and german language learners: a two-step process," in *LREC - 9th Language Resources and Evaluation Conference*, May 2014.

[11] K. Almeman, M. Lee, and A. A. Almiman, "Multi dialect arabic speech parallel corpora," in *Communications, Signal Processing, and their Applications (ICCSPA), 2013 1st International Conference on*, Feb 2013, pp. 1–6.

[12] TED, "Technology entertainment design," http://www.ted.com, 2006.

[13] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst, "Moses: open source toolkit for statistical machine translation," in *Proc. ACL 2007*, 2007, pp. 177–180.

[14] Raymond W. M. Ng, Mortaza Doulaty, Rama Doddipatla, Wilker Aziz, Kashif Shah, Oscar Saz, Madina Hasan, Ghada AlHarbi, Lucia Specia, and Thomas Hain, "The usfd slt system for iwslt 2014," in *Proc. IWSLT*, 2014.

[15] Pi-Chuan Chang, Michel Galley, and Christopher D. Manning, "Optimizing chinese word segmentation for machine translation performance," in *Proceedings of the Third Workshop on Statistical Machine Translation*, 2008, StatMT '08, pp. 224–232.

[16] B. Kingsbury, J. Cui, X. Cui, M. J. F. Gales, K. Knill, J. Mamou, L. Mangu, D. Nolden, M. Picheny, B. Ramabhadran, R. Schlter, A. Sethy, and P. C. Woodland, "A high-performance cantonese keyword search system," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 8277–8281.

[17] Z. Tske, D. Nolden, R. Schlter, and H. Ney, "Multilingual MRASTA features for low-resource keyword search and speech recognition systems," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 7854–7858.

[18] Tan Lee, W. K. Lo, P. C. China, and Helen Meng, "Spoken language resources for cantonese speech processing," *Speech Communication*, vol. 36, pp. 327–342, 2002.

[19] John Garofalo, David Graff, Doug Paul, and David Pallett, "CSR-I (WSJ0) complete," *Linguistic Data Consortium, Philadelphia*, 2007.

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, G. Ondrej, G. Nagendra, M. Hanneman, P. Motlicek, Q. Yanmin, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *Proceedings of the 2011 IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop*, Big Island, HA, 2011.