Siminchik: A Speech Corpus for Preservation of Southern Quechua

Ronald Cardenas[†], **Rodolfo Zevallos**[⋆], **Reynaldo Baquerizo**[‡], **Luis Camacho**[∓]
[†]*Institute of Formal and Applied Linguistics Charles University in Prague, Czech Republic*
[⋆‡∓]*Grupo de Telecomunicaciones Rurales Pontificia Universidad Católica del Perú*
[†]ronald.cardenas@matfyz.cz, [⋆]rjzevallos.salazar@gmail.com
[‡]baquerizo.reynaldo@pucp.edu.pe, [∓]camacho.l@pucp.pe

Abstract

Languages are disappearing at an alarming rate, linguistics rights of speakers of most of the 7000 languages are under risk. ICT play a key role for the preservation of endangered languages; as ultimate use of ICT, natural language processing must be highlighted since in this century the lack of such support hampers literacy acquisition as well as prevents the use of Internet and any electronic means. The first step is the building of resources for processing, therefore we introduce the first speech corpus of Southern Quechua, *Siminchik*, suitable for training and evaluating speech recognition systems. The corpus consists of 97 hours of spontaneous conversations recorded in radio programs in the Southern regions of Peru. The annotation task was carried out by native speakers from those regions using the unified written convention. We present initial experiments on speech recognition and language modeling and explain the challenges inherent to the nature and current status of this ancestral language.

Keywords: Quechua, endangered languages, corpus, speech recognition

1. Introduction

Peru is a multicultural country, not by modern immigration like USA or Europe but due the presence of many native first nations making a total of 10% of the population; the bulk of remainder has the same roots that these first nations but they lost ethnic identification decades or even centuries ago. First nations are speakers of 47 languages still alive but under risk of extinction.

Linguistic research of Peruvian native languages has achieved the creation of official standard alphabets of thirty five of these languages (Cáceres et al., 2016). Unfortunately, there is insufficient private and public funding to help increase, maintain and disseminate that knowledge. As a result, Peruvian native languages, amongst which Quechua is included, present scarce written footprint and even today they are predominantly orally transmitted. Even worse, the amount of digital content in Peruvian languages is extremely low, all of them are considered under-resourced, that means they meet one or more of the following characteristics:

1. they lack a unique writing system or an established grammar
2. they lack significant presence on Internet
3. they lack a critical mass of expert linguists
4. they lack electronic resources: monolingual corpora, bilingual electronic dictionaries, databases of transcribed speeches, pronunciation dictionaries or lexica.

Beyond linguistics, computational portability involves creating scalable language technology. However, language processing technologies require large datasets of joined and aligned text and speech, and parallel corpus. We can conclude that increasing the size of these kinds of corpora is an unavoidable task, despite the operational difficulties in doing so due to the lack of public funding.

In this paper, we present an initial effort to address computational portability for the Quechua family of languages. This way, we hope to boost the development of the complete set of software tools for this endangered language and finally to contribute to its revitalization.

Our main contribution is twofold:

- We present a 97 hour long speech corpus containing audio recordings of spontaneous conversations recorded in radio programs, for the two most widely spoken Quechua languages in Peru QUECHUA CHANCA and QUECHUA COLLAO.
- We perform experiments in traditional (acoustic-based) speech recognition and language modeling for this dataset.

2. Linguistics rights of Peruvian first nations

First nations have managed to survive colonial and post-colonial repression but are facing constant pressure from the side of the monolingual Spanish-speaking Peruvian society that drives them towards the abandonment of their ancestral languages.

A situation that is much less effectively addressed by programs for endangered languages is the case of historically important languages with large numbers of speakers but unclear official status, as Quechua. Such languages are located in regions where, for many centuries, their speakers have occupied an inferior socioeconomic position in trade with speakers of dominant languages; exposed to a strong economic dependence on the use of the dominant language, their feeling of pride of belonging to a language community has been hurt and now they are much more difficult to support than small speaker communities with a high degree of internal cohesion. Most of these languages are subject to a dramatic language shift that threatens to interrupt their transmission to future generations.(Adelaar, 2014).

The root of the problem is a ”structural inequality”, the legacy of colonialism is the inequality due to ethnicity. In South America, colonialism ended two hundred years ago but its consequences are still felt. In the last fifty years,

the Peruvian State issued laws to make compulsory the teaching of languages Quechua and Aymara and to declare Quechua as official language at national level (but later it was restricted just to the regional level). None of these laws really has been put into practice.
Issuance of laws is not enough if these are not enforced. All these good initiatives clashed with reality, a State apparatus neither convinced nor prepared. In spite of existing laws, in practice the State still ignores the multiculturalism and behaves as a monoculture and monolingual organization. Since this wrong paradigm is still in force, the State has not invested enough to build the linguistics skills to serve everyone equally. The consequences of this are the lack of promotion, discrimination and finally the isolation that leads to extinction of our native languages.
Top down decisions are essential but also is needed bottom up action from the grassroots. Our initiative is to change the wrong paradigm, to arouse national pride for our native roots, and to do so on three fronts:

1. demonstrate that our languages can be used in the modern technological world just like well established languages
2. demonstrate that our languages can carry contemporary culture and entertainment
3. demonstrate that our languages bring economic value to the nation, which justifies its preservation beyond the rights.

Nowadays social media should be accessible to everyone. People who are in some respect and/or to some extent functionally illiterate in a specific language are currently excluded from properly using these media. However, until now the high valuation of social media and any kind of ICT tool has reinforced the vicious circle of diglossia and even worse, created "cyberglossia". We expect turning the vicious circle into a virtuous circle by developing the three fronts mentioned, generating horizontal attitudes and practices towards all languages and reinforcing the potentialities and opportunities offered by technology related to endangered languages.
Quechua language is our first target. As said before, the basic standarization of this language is already done, the next steps to be undertaken are the building of corpus and the development of a lexical database and a spelling checker. In this paper we face the first of those challenges.

3. Background and Related Work

Even though there does not exist any Quechua speech dataset, to the best of our knowledge, various groups in Latin America and abroad have been working on Quechua language technology for the last few years. The Instituto de Lengua y Literatura Andina Amazónica (ILLA)[1] has been working on the construction of electronic dictionaries for Quechua, Aymara and Guaraní; the group Hinantin[2] at the Universidad Nacional San Antonio Abad del Cusco (UNSAAC) has produced a text-to-speech system for Cusco Quechua, a Quechua spell checker plug-in for LibreOffice (Rios, 2011) and a morphological analyzer for Ashaninka, an aboriginal language whose population is scattered across the Amazonian rainforest in Peru and Brazil.
Rios (Rios, 2016) describes a language technology toolkit that includes several things worth mentioning, such as the first morphological analyzer for Quechua, a hybrid machine translation in the direction Spanish-Quechua, and the first Quechua dependency treebank.
The AVENUE-Mapudungun project developed a machine translation system for Quechua and Mapudungun (Monson et al., 2006), and included textual and speech corpora. However, the speech dataset was only collected for Mapudungun. Resources and code can still be found in the website of the project Human Language Technology and the Democratization of information.[3]

4. The Quechua Language Familiy

Quechua is a family of languages spoken in South America with around 10 million speakers, not only in the Andean regions but also along the valleys and plains connecting the Amazonian Forest to the Pacific Ocean coastline.
Quechua languages are considered highly agglutinative with sentence structure subject-object-verb (SOV) and mostly post-positional. Table 1 contains an example of standard Quechua.

Quechua	Qichwa siminchik kan Qichwa simi-nchik ka-n
Lit. trans.	Quechua mouth-ours is
Translation	Quechua is our language.

Table 1: Sentence example of standard QUECHUA CHANCA

Even though the classification of Quechua languages remains open to research (Heggarty et al., 2005; Landerman, 1992), recent work in language technology for Quechua (Rios, 2016; Rios and Mamani, 2014) have adopted the categorization system described by Torero (Torero, 1964). This categorization divides the Quechua languages into two main branches, QI and QII. Branch QI corresponds to the dialects spoken in central Peru. QII is further divided in three branches, QIIA, QIIB and QIIC. QIIA groups the dialects spoken in Northern Peru, while QIIB the ones in Ecuador and Colombia. In this paper, we focus in the QIIC dialects, which correspond to the ones spoken in Southern Peru, Bolivia and Argentina. Mutual intelligibility between speakers of QI and QII dialects is not always given. However, QII dialects are close enough to allow mutual intelligibility (see Figure 1)
Siminchik gathers audio and text of two dialects from Southern Peru. The first one, QUECHUA CHANCA, is mainly spoken in Ayacucho and surrounding departments of Peru. The second one, QUECHUA COLLAO, is spoken in the departments of Cusco and Puno, and some Northern

[1] http://www.illa-a.org/wp/
[2] http://hinant.in
[3] https://code.google.com/archive/p/hltdi-l3/wikis/AvenueQuechuaCorpus.wiki

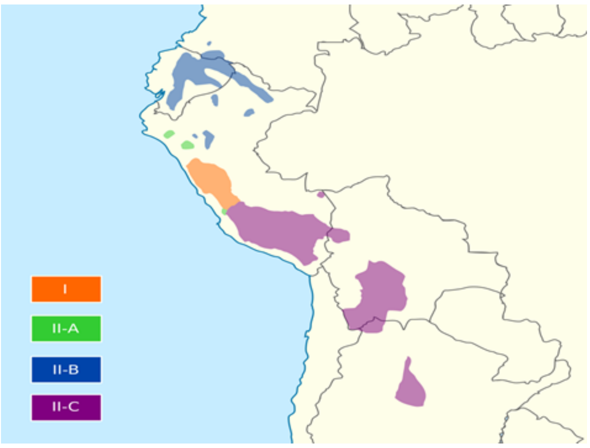

Figure 1: The four branches of Quechua language

	Bil	Alv	Pal	Vel	Uvu	Glo
Plosive	p	t	t͡ʃ	k	q	
Nasal	m	n	ɲ			
Fricative		s				h
Lat. Approx.		l	ʎ			
Approximant		ɹ				
Semi-consonants	w		y			

Table 2: Consonants in the phonemic inventory of QUECHUA CHANCA (IPA)

regions of Bolivia. The main difference between these dialects is the occurrence of glottalized and aspirated stops in QUECHUA COLLAO, a phonetic distinction that QUECHUA CHANCA lacks.

4.1. Normalization for Dialect Variation

While the semantic meaning and function of most of the 130 recognized morpheme units are shared by all languages in the family, the final forms (spoken and written) differs. In an effort to create a unified standard for written language, several standards were proposed, most notably the ones proposed by Profesor Cerron-Palomino (Cerrón-Palomino, 1994) and by the *Academy of Cusco Quechua*. In 1985, Ministry of Education of Peru issued the order 1218-85-ED setting the official alphabet of Quechua (and also of Aymara) which is still in force and this is the normalization system we reference as standard as well in the rest of this paper.
The standard proposed by the *Academy of Cusco Quechua* exclusively for QUECHUA COLLAO, presents a five-vowel system. Despite the fact this standard was rejected by the Ministry of Education and by linguists, because of the lack of formal and extensive study, *Academy of Cusco Quechua* keeps spreading its ideology (Coronel-Molina, 2007) causing interference in the dissemination of official standard.
It's worth to say that Cerron-Palomino's standardized orthography was adopted by the Bolivian government with only one change: the glottal fricative consonant is written as j instead of h.

4.2. Phonemic Inventory

QUECHUA CHANCA has a total of 15 consonants, most of them voiceless, as shown in Table 2. As in Spanish, the phoneme [t͡ʃ] is written as *ch*, [ɲ] as *ñ* , and [ʎ] as *ll*.

QUECHUA COLLAO also has a glottal and an aspirated version of each plosive consonant, leading to a total of 25 consonants. However, the use of voiced consonants present in the Spanish phonemic inventory is common due to the large number of borrowings present in all dialects.
Both quechua dialects present only the vowels *a* [æ], *i* [ɪ] and *u* [ʊ], although in proximity of /q/, they are pronounced as [ɑ], [ɛ] and [ɔ], respectively. The online database OMNIGLOT has entries for both the Chanca [4] and the Collao standard [5].

5. Siminchik Speech Corpus

The collection and curation of SIMINCHIK (Quechua word for *"our language"*) speech corpus is part of a much bigger initiative that seeks to digitally preserve ancestral Quechua and revitalize it for daily use by current and future generations. Such a colossal enterprise requires thousands of hours of quality audio from every dialect along with their transcriptions.

5.1. Crowd-Sourcing Transcription Annotation

In our effort to construct an audio corpus and a corresponding transcribed text corpus, we first began collecting audio recorded from radio programs. These included QUECHUA CHANKA from Ayacucho, Apurimac and QUECHUA COLLAO from Cusco, Puno.
The initial collection comprised of hundreds of raw audio hours that contained spot advertisements, music and Spanish speech. To separate speech from music we employed a voice detector using pyAudioAnalysis (Giannakopoulos, 2015). The audio was converted to mono channel, resampled at 16 kHz, encoding with 16 bits precision and saved in the WAV format. In preliminary experiments, it was found that radio locutors speak fast enough for the voice detector to not be able to locate pauses between sentences. Hence, we opted for splitting any audio detected as voice in segments no longer than 30 seconds each. The main reason for this was to avoid dealing with long sequences, as it is known that sequence models' performance decay rapidly with the length of the sequence. This resulted in words quite possibly being truncated at the beginning and at the end of the segment.
Given the large amount of data, we crowdsourced the task of annotating orthographic transcriptions for each audio segment. We built a specialized website for this purpose [6] and called for native speakers of both dialects to voluntarily participate. Each audio clip was annotated by exactly two annotators. The number of annotators reached and transcribed hours are presented in Table 3.
The website addressed the accuracy limitations of the voice detector by allowing the users to mark a given audio as containing only Spanish speech, only music, Quechua speech

[4] https://www.omniglot.com/writing/ayacuchoquechua.htm
[5] https://www.omniglot.com/writing/quechua.htm
[6] https://siminchikkunarayku.pe/corpus/transcribir

	Female		Male	
	Quechua Chanca	Quechua Collao	Quechua Chanca	Quechua Collao
# annotators	160	63	153	56
# transcribed hours	58.28	12.95	17.74	8.03

Table 3: Details of volunteer annotators

(possibly with non-loud music in the background) or entirely unintelligible. The interface also includes the option to indicate in which second the transcription begins and ends. This is in order to make up for a potentially truncated initial or final word because of the audio's hard splitting.

The statistics of the speech corpus are presented in Table 4. A total of 97 hours have been transcribed so far.

5.2. Text Corpus

Although the collected speech data was sufficient for acoustic modeling, it was insufficient for n-gram estimation. Hence, we gathered text corpora from legal documents, such as the Peruvian Constitution and the Declaration of Human Rights; as well as Andean tales for children prepared by the Ministry of Education of Peru, and the New Testament of the Bible. The text is normalized for further downstream usage using (Rios, 2016)' normalizer.

The statistics of the final text corpus are shown in Table 5. For our language modeling experiments, this corpus was split into three corpora for training, evaluation (validation) and testing, in an 80/10/10 ratio.

5.3. Preprocessing and Normalization of Transcriptions

Besides standard preprocessing for speech recognition, such as punctuation removal and case normalization, special care had to be taken for the normalization of interjections, those common to all dialects and those rooted in the culture of the community of the speaker. For this purpose we created a dictionary of said interjections (e.g. *mamay, ayy, mmm, ahaa*) and crafted regular expressions in order to map each orthographic variation to a fixed word form. Annotators were encouraged to use the Spanish writing system for proper names, since almost all of these are loan words from the Spanish language.

After cleaning the text, it was necessary to normalize dialectal variations to the standard written format. We used the morphological analyzer and normalizer developed by (Rios, 2016), which normalizes to the CHANCA standard. However, spelling error still remained unchanged, since annotators were not provided a spell checker in the web tool. We dealt with this by using a language model to score the two transcriptions of each audio segment and choosing the most probable one. By doing this, we keep only the less noisy transcription for each audio segment. This language model was trained over the normalized corpus presented in section 5.2.. Section 6.2. presents details of the design and experimental results of this model. Then, we manually curated all remaining transcriptions.

6. Experiments

In this section we analyze the quality of our speech corpus by conducting acoustic and language modeling experiments.

6.1. Acoustic Modeling

The acoustic model for SIMINCHIK was built using the Hidden Markov Model Toolkit (HTK) (Young et al., 2002) because it has a better performance in dynamic decoder (dictionary, language model and acoustic model) (Ganesh and Sahu, 2015) and phonetic segmentation with a tolerance region less than 30ms (Matoušek and Klíma, 2017).

We sub-sampled the dataset, obtained 8 hours of training and 2 hours of testing data for a total of 16,340 instances. The training set includes audio with 9 speakers (3 male and 6 female), while the test set includes two (one male and one female). We work just with 10 out of a total of 97 hours of audio because only this 10 hours were speech-text aligned at word level, as required to run the experiment; the 87 hours left were aligned at sentence level.

Each utterance was split into Hamming windows of 25ms with an offset of 10ms. Acoustic parameters were 39 MFCCs with 12 Mel cepstrum plus log energy and their delta (first order derivative) and acceleration (second order derivative) coefficients.

Our monophone models consist of 5-state HMMs in which the first and last state are non-emitting states. The prototype model were initialized using HCompV which allows for a fast and precise convergence of the training algorithm. Next, HMM parameters were re-estimated 5 times using the HERest and HHEd which computes the optimum values of HMM parameters using gaussian mixtures. These parameters were re-estimated repeatedly using the training data until re-estimation converged.

The experiment follow the procedures mentioned in (Odriozola et al., 2014; Dua et al., 2012).

Table 6 shows promising results for word level accuracy and word error rate, obtaining averages of 82.8% and 17.2%, respectively.

6.2. Language Modeling

Given the high presence of rare words in morphologically rich languages such as Quechua, we use a singleton pruning rate κ of 0.05 as proposed by (Botha, 2015), in order to randomly replace only a fraction κ of words occurring only once in the training data with a global UNK symbol.

We built a 5-gram interpolated Modified Kneser-Ney model at the word level, obtaining a perplexity of 298.79. A vocabulary intersection analysis further revealed a 63.46% intersection between the validation set's and training set's vocabulary. This rather high value of perplexity and low vo-

	Total	Training	Validation	Test
Words	448,919	404,625	22,526	21,768
Vocabulary	89,767	83,258	9,538	9,433
Length	97.5h	88h	4.8h	4.7h

Table 4: Statistics of the Quechua dataset.

	Total	Training	Validation	Test
Words	1'211,936	1'011,335	103,478	97,123
Vocabulary	174,151	155,655	30,776	29,024
Sentences	70,790	57,735	6,300	6,755

Table 5: Text corpus statistics

cabulary intersection reveal the morphological richness of this family of languages. A more thorough inspection of the vocabulary showed that 64.42% of words have a frequency of one.

Speaker	# utterances	WACC	WER
S1	1,509	83.2	16.8
S2	3,000	82.2	17.8
Total	4,509	82.8	17.2

Table 6: Acoustic model results for two speakers (one male and one female) in per-word accuracy (WACC) and word error rate (WER)

7. Conclusions & Further Work

Driven by social and attitudinal factors deeply rooted in history, massive language shift of endangered languages with millions of speakers is difficult to reverse, and it requires more complex techniques than language maintenance and revitalization in small language communities.

By the fact that Quechua official spelling hasn't been well disseminated, Quechua speakers are to some extent functionally illiterate in their own language. Our approach will in the end enable them to use spoken input, that will be converted automatically in standard written format.

As first step of that medium-term goal, we have collected and curated raw audio from radio shows to build the first dataset of Quechua speech: SIMINCHIK. It consists of 97 hours of spontaneous Quechua speech and the corresponding transcribed text for the Chanca and Collao, dialects from the southern regions of Peru.

We present initial ASR experiments using a HMM acoustic model, yielding a promising word error rate of 17.2%. The highly agglutinative nature of the language is challenging for word form based tasks, such as n-gram language modeling, POS tagging, speech recognition, among others. Similar challenges can be found in ASR systems for typologically similar languages such as Turkish (Carki et al., 2000) and Basque (Odriozola et al., 2014).

We are currently introducing neural networks instead a HMM acoustic model. The new model is to reproduce the work of (Graves and Jaitly, 2014) including extensions made by (Amodei et al., 2016), that means to introduce neural networks in each stage of the process.

Further work on ASR of Quechua language will proceed in use Knowledge Transfer described by (Zhao et al., 2014) to build a model capable of learning several languages, in this case Basque and Quechua.

Finally, this research is a part of Siminchikkunarayku [7] initiative which vision is that the future of the languages of America depends upon the conservation of the native languages, but also upon the polyglotism of the citizens. Siminchikkunarayku follows two orientations:

1. "language as a right" views minority languages as a right to which their speakers are entitled. Following this, preservation of Quechua is an identity reinforcement and recognition of that population.

2. "language as a resource" sees the minority languages as potential economical resources for the whole nation. A lot of research must be done to make visible the profitability of preservation.

8. Acknowledgements

This project was supported by CONCYTEC CIENCIACTIVA of the Peruvian government through grant 164-2015-FONDECYT and by PUCP through grant 2017-3-0039/436

9. Bibliographical References

Adelaar, W. F. (2014). Endangered languages with millions of speakers: Focus on quechua in peru. *JournaLIPP*, (3):1–12.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al. (2016). Deep speech 2: End-to-end speech recognition in english and mandarin. In *International Conference on Machine Learning*, pages 173–182.

Botha, J. A. (2015). Probabilistic modelling of morphologically rich languages. *arXiv preprint arXiv:1508.04271*.

Cáceres, R., Cavero Cornejo, O., Gutiérrez, D., et al. (2016). Diagnóstico descriptivo de la situación de los pueblos originarios y de la política de educación intercultural bilingüe en el perú.

[7] https://siminchikkunarayku.pe

Carki, K., Geutner, P., and Schultz, T. (2000). Turkish lvcsr: towards better speech recognition for agglutinative languages. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 3, pages 1563–1566. IEEE.

Cerrón-Palomino, R. (1994). Quechua sureño diccionario unificada quechua-castellano castellano-quechua [unified dictionary of southern quechua, quechua-spanish spanish-quechua]. *Lima: Biblioteca Nacional del Perú*.

Coronel-Molina, S. M. (2007). *Language policy and planning, and language ideologies in Peru: The case of Cuzco's High Academy of the Quechua Language (Qheswa simi hamut'ana kuraq suntur)*. University of Pennsylvania.

Dua, M., Aggarwal, R., Kadyan, V., and Dua, S. (2012). Punjabi automatic speech recognition using htk. *IJCSI International Journal of Computer Science Issues*, 9(4):1694–0814.

Ganesh, D. S. and Sahu, P. K. (2015). A study on automatic speech recognition toolkits. In *International Conference on Microwave, Optical and Communcation Engineering*.

Giannakopoulos, T. (2015). pyaudioanalysis: An open-source python library for audio signal analysis. *PloS one*, 10(12).

Graves, A. and Jaitly, N. (2014). Towards end-to-end speech recognition with recurrent neural networks. In *International Conference on Machine Learning*, pages 1764–1772.

Heggarty, P., Valko, M. L., Huarcaya, S. M., Jerez, O., Pilares, G., Paz, E. P., Noli, E., and Usandizaga, H. (2005). Enigmas en el origen de las lenguas andinas: aplicando nuevas técnicas a las incógnitas por resolver. *Revista Andina*, 40:9–57.

Landerman, P. N. (1992). Quechua dialects and their classification. *PhD Thesis*.

Matoušek, J. and Klíma, M. (2017). Automatic phonetic segmentation using the kaldi toolkit. In *International Conference on Text, Speech, and Dialogue*, pages 138–146. Springer.

Monson, C., Llitjós, A. F., Aranovich, R., Levin, L., Brown, R., Peterson, E., Carbonell, J., and Lavie, A. (2006). Building nlp systems for two resource-scarce indigenous languages: Mapudungun and quechua. *Strategies for developing machine translation for minority languages*, page 15.

Odriozola, I., Serrano, L., Hernaez, I., and Navas, E. (2014). The ahosr automatic speech recognition system. In *Advances in Speech and Language Technologies for Iberian Languages*, pages 279–288. Springer.

Rios, A. and Mamani, R. C. (2014). Morphological disambiguation and text normalization for southern quechua varieties. *COLING 2014*, page 39.

Rios, A. (2011). Spell checking an agglutinative language: Quechua. *University of Zurich. Zurich Open Repository and Archive*.

Rios, A. (2016). A basic language technology toolkit for quechua.

Torero, A. (1964). Los dialectos quechua.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., et al. (2002). The htk book. *Cambridge university engineering department*, 3:175.

Zhao, Y., Xu, Y. M., Sun, M. J., Xu, X. N., Wang, H., Yang, G. S., and Ji, Q. (2014). Cross-language transfer speech recognition using deep learning. In *Control & Automation (ICCA), 11th IEEE International Conference on*, pages 1422–1426. IEEE.