PHONETIC LESSONS FROM AUTOMATIC PHONEMIC TRANSCRIPTION: PRELIMINARY REFLECTIONS ON NA (SINO-TIBETAN) AND TSUUT'INA (DENE) DATA

Alexis Michaud* Oliver Adams** Christopher Cox*** Séverine Guillaume*

*Langues et Civilisations à Tradition Orale, UMR 7107 CNRS / Sorbonne Nouvelle
**Center for Language and Speech Processing at Johns Hopkins University
***School of Linguistics and Language Studies at Carleton University
alexis.michaud@cnrs.fr oliver.adams@gmail.com christopher.cox@carleton.ca severine.guillaume@cnrs.fr

ABSTRACT

Automatic phonemic transcription tools now reach high levels of accuracy on a single speaker with relatively small amounts of training data: on the order of 100 to 250 minutes of transcribed speech. Beyond its practical usefulness for language documentation, use of automatic transcription also yields some insights for phoneticians. The present report illustrates this by going into qualitative error analysis on two test cases, Yongning Na (Sino-Tibetan) and Tsuut'ina (Dene). Among other benefits, error analysis allows for a renewed exploration of phonetic detail: examining the output of phonemic transcription software compared with spectrographic and aural evidence. From a methodological point of view, the present report is intended as a case study in Computational Language Documentation: an interdisciplinary approach that associates fieldworkers ("diversity linguists") and computer scientists with phoneticians/phonologists.

Keywords: speech recognition, machine learning, error analysis, interdisciplinarity, Computational Language Documentation.

1. INTRODUCTION: PHONETICS AND AUTOMATIC SPEECH RECOGNITION

Speech recognition has progressed in recent years, but with less collaboration between computer scientists and linguists than one could wish for: improved performance is mostly gained by leveraging the power of new statistical tools and new hardware. A lot is nonetheless at stake in collaborations between linguists and specialists of Natural Language Processing. Hand-crafted features can be meaningfully integrated in deep learning [23], with more promising results than under an 'end-to-end' black-box approach (see also [10]). Interdisciplinary dialogue is as relevant in the age of machine learning as ever, and it can be argued that phoneticians are especially well-prepared for interdisciplinary work because phonetics is a highly interdisciplinary field, with strong ties to acoustics, physiology and computer modelling as well as to the humanities. This point is illustrated here by reporting on lessons learnt when using an automatic phoneme transcription tool, `Persephone` (/**pər**ˈ**sɛfəni**/) [1], which can build an effective single-speaker acoustic model on the basis of limited training data, on the order of 100 to 250 minutes of transcribed speech. Emphasis is not placed on the tool's practical usefulness for language documentation [6, 18], but on opportunities that it offers for phonetic research.

2. METHOD

2.1. The automatic phonemic transcription method

The phonemic transcription tool used in this study implements a *connectionist temporal classification* model similar to that of [7]. Filterbank features (analogous to a spectrogram) are extracted from the waveform in overlapping 10ms frames. These are then fed into a multi-layer *recurrent neural network* which predicts the probability of a transcription symbol given the frame and its surrounding acoustic context. A key feature of this approach is that the model has no hard notion of phoneme or segment boundaries. Suprasegmental acoustic information can be captured by the neural network, which has the capacity to make predictions based on both immediate and long-ranging acoustic cues. The code and a link to documentation can be found at https://github.com/persephone-tools/persephone.

2.2. Cross-validation: creating 'parallel-text' versions to compare the linguist's transcription with an automatically generated transcript

To compare manual transcripts with automatically generated transcripts, one of the transcribed texts is set aside, and an acoustic model is trained on the rest

of the corpus (the *training set*), then applied to the target text. This procedure, referred to technically as "cross-validation", was applied to each of the texts in turn.

2.3. Choice of qualitative analysis

It is common to perform quantitative evaluation of such models against a human reference transcription using phoneme error rate, as was done for Yongning Na in [1]. In this study, we complement such quantitative investigations with qualitative analysis of the errors. To facilitate this, we generated parallel-text files (in PDF format) with colour-coded inconsistencies between the manual transcripts and the automatically generated transcripts, which we then hone in on for qualitative analysis.

3. YONGNING NA: HIGHLIGHTING THE ACOUSTIC SPECIFICITY OF LONG WORDS

Yongning Na is a Sino-Tibetan language of Southwest China [16]. A repository dedicated to Na data has a specific folder for materials related to `Persephone`: https://github.com/alexis-michaud/na/tree/master/Persephone. The complete set of 'parallel-text' versions of the twenty-seven Na narratives available to date is available in the folder `2018_08_StoryFoldCrossValidation`. The various other materials of the present study (including the manual transcriptions) are also available for download from the same repository, following principles of Open Science (as advocated e.g. by [3]). The audio files with annotated transcriptions can be consulted online in the Pangloss Collection [14].

3.1. High error rate on a quadrisyllabic proper name: qualitative observations

An example of the parallel-text view is shown below, with highlighted differences between the linguist's transcription, in the first line, and the automatic transcription (the acoustic model's best hypothesis) in the second line. Glosses are provided in (1).

ɻ̩˩ ʈʂʰe˦ ɖɯ˩ mɑ˩ ɻ̩˩ ʈʂʰe˦ ɖɯ˩ mɑ˩ pi˦ dʑo˩
æ̃˩ ʈʂʰe˦ ɖɯ˦ mæ˩ ʈʂʰɯ˦ bi˦ mæ˩ pi˦ dʑo˩

(1) ɻ̩˩ʈʂʰe˦ ɖɯ˩mɑ˩ pi˦ -dʑo˩
Erchei_Ddeema to_say TOP
" 'Erchei Ddeema! Erchei Ddeema!' [she] called out."

(*BuriedAlive2*, sentence 13; direct access: https://doi.org/10.24397/pangloss-0004537#S13)

In this short example, transcription errors occur on the two occurrences of the proper name 'Erchei Ddeema' (the name of one of the characters in the story), phonemically /ɻ̩˩ʈʂʰe˦ ɖɯ˩mɑ˩/. In view of the tool's overall low phoneme error rate (on the order of 17%), it is striking to find nine errors on tones and consonants in a sequence of just eight syllables. Examining all occurrences of this name in the text, it turned out that none was devoid of mistakes: see Table 1.

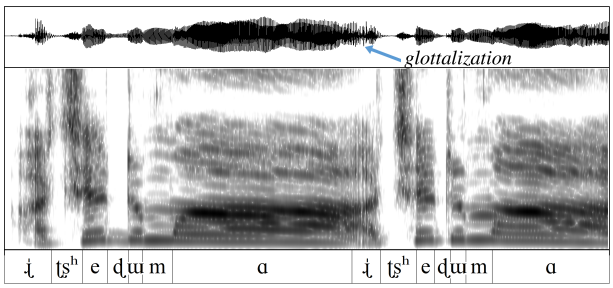
Table 1: Automatic transcription of the eleven instances of the name 'Erchei Ddeema', /ɻ̩˩ʈʂʰe˦ ɖɯ˩mɑ˩/, occurring in the narrative *BuriedAlive2*.

sentence	syll. 1	syll. 2	syll. 3	syll. 4
S13	p æ ˩	ʈʂʰ ɯ ˥	ɖ ɯ ˦	m ɻ̩ ˩
S14 (1st)	æ̃ ˩	ʈʂʰ e ˦	ɖ ɯ ˦	m æ ˩
S14 (2nd)		ʈʂʰ ɯ ˦	b i ˦	m æ ˩
S18	ɑ ˩	ʈʂʰ e ˦	ɖ ɯ ˦	m ɤ ˩
S77		tʰ i ˥	ɖ ɯ ˦	m ɑ ˩
S105	æ ˩	ʈʂʰ ɯ ˦	ɖ ɯ ˦	m ɤ ˦
S106	ɻ̩ ˩	ʈʂʰ e ˦	ɖ ɯ ˦	m ɤ ˦
S107	ɻ̩ ˦	ʈʂʰ ɯ ˦	dʑ ɯ ˦	m ɤ ˦
S129	æ ˩	ʈʂʰ ɯ ˦	ɖ ɯ ˦	m ɤ ˦
S132	ɻ̩ ˩	ʈʂʰ ɯ ˦	ɖ ɯ ˦	m ɤ ˦
S147	æ ˩	ʈʂʰ ɯ ˦	ɖ ɯ ˦	m ɤ ˦
reference	ɻ̩ ˩	ʈʂʰ e ˦	ɖ ɯ ˩	m ɑ ˩

The first syllable, a syllabic approximant /ɻ̩/, is identified as a vowel in six cases, i.e. there is not enough acoustic evidence of retroflexion for identification as /ɻ̩/. (The initial **p** in S13 is not a surprising mistake: a hard onset – initial glottal stop – can be difficult to distinguish acoustically from **p**.) This syllable goes unnoticed in two examples where it follows the preceding vowel without a sharp acoustic discontinuity. Fig. 1 shows a spectrogram. A brief glottalized span is visible; it presumably contributes to signalling phrasing, as in various other languages [11, p. 3218]. This glottalization may be in part responsible for lack of detection of the [ɻ̩] despite the presence of hints such as a final decrease in the third formant.

The vowel in the second syllable is identified as /ɯ/ in a majority of cases. In Na, /ɯ/ has an apical allophone after retroflex fricatives and affricates, i.e. /ʈʂʰɯ/ is realized as [ʈʂʰʐ̩]. Classification as /ɯ/ instead of /e/ can therefore be interpreted as a case of hypo-articulation of the vowel: the tongue's movement towards a [e] target is not as ample as in the statistically dominant pattern (as identified by the automatic transcription software). The tongue remains close to the configuration that it adopted for the consonant [ʈʂʰ], leading to the identification of

Figure 1: Example S14, where the second occurrence of /ʑ̩/ was not identified by automatic transcription software. (Scale of spectrogram: 0-8,000 Hz. Time scale: 1.78 second.)



the syllable as [**tʂʰz̩**] (phonemically /**tʂʰɯ**/). Categorization of the vowel of the fourth syllable as /**ɤ**/ instead of /**ɑ**/ is also interpreted as resulting from hypo-articulation.

The third syllable is least affected by misidentification, but its tone is systematically identified as Mid (˧) instead of Low (˩). Acoustically, the quadrisyllabic name's /L.M.L.L/ pattern is realized with higher f_0 values on the middle syllables (the third as well as the second) and somewhat lower f_0 values on the first and last syllables. This is reminiscent of word-level patterns found in polysyllabic languages, a similarity which allows us to proceed to an interpretation.

3.2. Interpretation of the findings

In Na, lexical roots are monosyllabic, following dramatic phonological erosion in the course of history [9]. These roots combine anew into disyllables through compounding and affixation, so that disyllables are widely attested, and combine, in turn, into longer words [15]. Words of four syllables or more make up about 6% of a 3,000-word lexicon [17] and their frequency of occurrence in the 27 texts is similar (5.5%). Quadrisyllables are thus marginal. This fact is held to be key to the errors shown in Table 1: the acoustic model tends to 'overfit' to the statistically common types (monosyllabic or disyllabic morphemes, with limited phonological material, and consequently articulated with precision), to the detriment of the less common types (long words, with enough phonological materials that some can be hypo-articulated with little threat to intelligibility). It should not come as a surprise to phoneticians with an interest in the typology of word structures and prosodic structures. But analysis of automatically generated transcriptions opens fresh perspectives for investigating the *hierarchy* of factors influencing allophonic variation. These factors are known to include the nature of the words (lexical words vs. function words); the extent to which function words are 'hypo-articulated' (weakened) varies across languages [4]. In Na, there is no conspicuous difference between function words and lexical words in terms of error rates in phonemic recognition; this observation (which remains to be quantified) suggests that the acoustic difference is relatively limited, in comparison with acoustic differences between words of different *lengths*. There is thus a hope of gaining typological insights into differences across languages in the relative importance of the various factors that contribute to allophonic variation.

From a practical point of view, these findings suggest that gains in accuracy can be obtained in future work by incorporating word boundary information in the training set. This information is present in the original data set, but had been removed when preprocessing the data to serve as training corpus.

4. TSUUT'INA (DENE): REVEALING THE PHONEMIC VALIDITY OF AN ORTHOGRAPHIC CONTRAST

Tsuut'ina, a Dene language spoken in southern Alberta, Canada, has been analyzed in previous descriptive linguistic studies as having four phonemic vowels, *i, a, o, u* (IPA: /**ɪ a ɒ ʊ**/) [13, 5]. Recent acoustic studies based on the speech of several first-language Tsuut'ina speakers have arrived at different conclusions as to the synchronic reality of this analysis, however, particularly concerning the independence of the low vowels /**a**/ and /**ɒ**/. Whereas [2] concludes that these four vowels are still distinct phonemes, even though the acoustic distance between /**a**/ and /**ɒ**/ is small, [20, 21] hypothesizes that these differences are instead allophonic realizations of a single low vowel phoneme (i.e., /**ɒ**/ when long or when short and appearing before a back consonant, and /**a**/ elsewhere). The close relationship that exists between these two vowels is reflected in metalinguistic observations of first-language speakers of Tsuut'ina, as well, who have reported that the apparent contrast between /**a**/ and /**ɒ**/ may be marginal and not consistently realized in all cases or by all speakers where it would be expected historically.

In the materials used in the present study, it thus was an open question whether or not this contrast was consistently present in a recognizable way in spontaneous speech. The materials used as a training set to create an acoustic model with `Persephone` nonetheless contain the distinction between /**a**/ and /**ɒ**/ (orthographic *a* and *o*), because the model trained for Tsuut'ina takes as input an orthographic repre-

sentation that is phonemic in orientation, not a string of IPA symbols. This presents an opportunity to explore the nature of this distinction further: if no consistent phonetic contrast was being made at all between /**a**/ and /**ɒ**/, then the statistical model would not be able to distinguish them consistently.

Interestingly, the acoustic model does a surprisingly good job of distinguishing the two hypothesized phonemes, /**a**/ and /**ɒ**/, both in short and long realizations. This can be interpreted as evidence that the speaker of Tsuut'ina represented in the audio materials still makes the distinction. It is not at all impossible that acoustic models could outperform linguists that are non-native speakers of the language they work on and who have difficulty hearing certain phonemic contrasts.

A methodological caveat is in order here. As explained in §2.1, the acoustic model takes context into account. It is theoretically possible that the neural network learnt to distinguish where to transcribe /**a**/ and /**ɒ**/ on the basis of context by using the text in the training data, rather than acoustic characteristics found in the signal. Thankfully, there is some evidence that speaks against that possibility here. If this contrast were allophonic and predictable from consonantal environment and vowel length, then one might expect the statistical model to struggle to distinguish /**a**/ and /**ɒ**/ in long vowels, where any phonetic contrast is reportedly neutralized [20]. But in the Tsuut'ina materials considered here, we find little conclusive evidence to suggest that segmental context is sufficient on its own to allow for the level of consistency in disambiguation that automatic phoneme recognition provides. In the case of long vowels, where phonetically grounded distinctions between these segments are reportedly minimized, we find /**a**/ and /**ɒ**/ appearing in the same segmental environments (e.g., word-finally after **k'**, as in *k'oo* /**k'ɒː**/ 'recent' vs. *chák'aa* /**tʃʰák'aː**/ 'rib').

While this suggests that context alone cannot fully distinguish these vowels, it is still possible that the model may be learning the statistical distribution of these phonemes over segmental contexts from the textual training data and applying that information to the task of recognition (e.g., all other things being equal, returning /**ɒ**/ rather than /**a**/ in specific contexts where the former vowel is more frequent than the latter). The extent to which it does this has not been rigorously explored yet, and merits further attention in future research.

Use of automatic phonemic transcription thus has the side benefit of offering additional evidence on a difficult aspect of the Tsuut'ina phonemic system. The support that automatic phonemic transcription systems such as `Persephone` offer in distinguishing phonetically close segments such as these has further proven valuable in preparing new transcriptions of Tsuut'ina materials, providing phonetically grounded hypotheses against which the intuitions of contributing speakers and transcribers can be compared.

5. FUTURE WORK

The work reported here is still at an early stage. There are many topics to explore, and there is much testing to be done. One point that seems worth highlighting is the perspective of extracting knowledge from the acoustic model: attempting to retrieve knowledge from the acoustic models generated through machine learning. Machines follow procedures that differ from those of linguists, and reflections on these (statistical) procedures could help characterize phonemes in terms of their defining acoustic properties, going beyond the categorization allowed by the International Phonetic Alphabet symbols and diacritics: for instance, characterizing differences between phonemes transcribed as /**i**/ in different languages [22]. Varying the input and evaluating differences in the output (i.e. conducting *ablation studies*) is one way to assess the role of different types of information in the acoustic signal.

Due to the nature of the statistical models, known as 'artificial neural-network models,' it is not easy to retrieve knowledge from the model: spelling out which acoustic properties are associated with which phonemes. Software based on a neural-network architecture is generally used as a black box. But there is a growing area of research on devising methods to open the box in order to relate what the model predicts (in the case of `Persephone`: the phonemes, tones, and tone-group boundaries) to input variables that are readily interpretable, and which humans can make sense of [19, 8, 12]. The use of such methods has the potential to amplify the insights the tool of speech recognition technology can provide to the phonetic sciences.

6. CONCLUSION

The insights presented above constitute a side benefit of team work in the budding interdisciplinary field of Computational Language Documentation. One discipline's 'by-products' can constitute relevant input for another, and what constitutes mere *application* in one field (such as development of a fine-tuned phonemic transcription tool) can open new research perspectives in another field.

7. ACKNOWLEDGMENTS

Many thanks to the Na and Tsuut'ina language consultants, colleagues, and friends, including the Office of the Tsuut'ina Language Commissioner for support of this work. We gratefully acknowledge financial support for the development of `Persephone` from the University of Queensland and from a Transdisciplinary Innovation Grant of the Australian Research Council Centre of Excellence for the Dynamics of Language. This work is a contribution to the "Empirical Foundations of Linguistics" Labex project (ANR-10-LABX-0083).

8. REFERENCES

[1] Adams, O., Cohn, T., Neubig, G., Cruz, H., Bird, S., Michaud, A. 2018. Evaluating phonemic transcription of low-resource tonal languages for language documentation. *Proceedings of LREC 2018 (Language Resources and Evaluation Conference)* 3356–3365.

[2] Barreda, S. 2011. The Tsuut'ina vocalic system. *Rochester Working Papers in the Language Sciences* 6, 1–10.

[3] Berez-Kroeker, A. L., Gawne, L., Kung, S. S., Kelly, B. F., Heston, T., Holton, G., Pulsifer, P., Beaver, D. I., Chelliah, S., Dubinsky, S., Meier, R. P., Thieberger, N., Rice, K., Woodbury, A. C. 2018. Reproducible research in linguistics: a position statement on data citation and attribution in our field. *Linguistics* 56(1), 1–18.

[4] Brunelle, M., Chow, D., Nguyễn, T. N. U. 2015. Effects of lexical frequency and lexical category on the duration of Vietnamese syllables. *Proceedings of the 18th International Congress of Phonetic Sciences* Glasgow. 1–5.

[5] Cook, E.-D. 1984. *A Sarcee grammar*. University of British Columbia Press.

[6] Foley, B., Arnold, J., Coto-Solano, R., Durantin, G., Ellison, T. M. 2018. Building speech recognition systems for language documentation: the CoEDL Endangered Language Pipeline and Inference System (ELPIS). *Proceedings of the 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU), 29-31 August 2018* Gurugram, India. ISCA 200–204.

[7] Graves, A., Mohamed, A., Hinton, G. May 2013. Speech recognition with deep recurrent neural networks. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* 6645--6649.

[8] Hohman, F., Head, A., Caruana, R., DeLine, R., Drucker, S. M. 2019. Gamut: a design probe to understand how data scientists understand Machine Learning models. *Proceedings of ACM CHI Conference on Human Factors in Computing Systems* Glasgow.

[9] Jacques, G., Michaud, A. 2011. Approaching the historical phonology of three highly eroded Sino-Tibetan languages: Naxi, Na and Laze. *Diachronica* 28(4), 468–498.

[10] Jimerson, R., Prud'hommeaux, E. 2018. ASR for documenting acutely under-resourced indigenous languages. *Proceedings of LREC 2018 (Language Resources and Evaluation Conference)* Miyazaki. 4161–4166.

[11] Kuang, J. 2017. Creaky voice as a function of tonal categories and prosodic boundaries. *Proceedings of Interspeech 2017* Stockholm. 3216–3220.

[12] Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.-R. 2019. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature communications* 10(1), 1096.

[13] Li, F.-K. 1930. A study of Sarcee verb-stems. *International Journal of American Linguistics* 6(1), 3–27.

[14] Michailovsky, B., Mazaudon, M., Michaud, A., Guillaume, S., François, A., Adamou, E. 2014. Documenting and researching endangered languages: The Pangloss Collection. *Language Documentation and Conservation* 8, 119–135.

[15] Michaud, A. 2012. Monosyllabicization: Patterns of evolution in Asian languages. In: Nau, N., Stolz, T., Stroh, C., (eds), *Monosyllables: From phonology to typology*. Berlin: Akademie Verlag 115–130.

[16] Michaud, A. 2017. *Tone in Yongning Na: lexical tones and morphotonology*. Berlin: Language Science Press.

[17] Michaud, A. 2018. *Na (Mosuo)-English-Chinese dictionary*. Paris: Lexica.

[18] Michaud, A., Adams, O., Cohn, T., Neubig, G., Guillaume, S. 2018. Integrating automatic transcription into the language documentation workflow: experiments with Na data and the Persephone toolkit. *Language Documentation and Conservation* 12, 393–429.

[19] Montavon, G., Samek, W., Müller, K.-R. 2017. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* 73, 1–15.

[20] Sims, M. 2010. Vowel space in the Athabaskan language of Tsuut'ina: Vowel shift in language attrition. *Proceedings of the Conference on the Endangered Languages and Cultures of Native America* volume 1.

[21] Sims, M. 2011. Acousting phonetic analysis as a means of defining the phonemic inventory: Evidence from the vowel space of Tsuut'ina. *Rochester Working Papers in the Language Sciences* 6, 1–18.

[22] Vaissière, J. 2011. On the acoustic and perceptual characterization of reference vowels in a cross-language perspective. *Proceedings of the 17th International Congress of Phonetic Sciences* Hong Kong.

[23] Wu, M., Liu, F., Cohn, T. 2018. Natural language processing not-at-all from scratch: Evaluating the utility of hand-crafted features in sequence labelling. *2018 Conference on Empirical Methods in Natural Language Processing* Brussels.