Phonetic lessons from automatic phonemic transcription: preliminary reflections on Na (Sino-Tibetan) and Tsuut'ina (Athabaskan) data

Alexis Michaud, Oliver Adams, Christopher Cox, Séverine Guillaume

Non-final version. To appear in: Proceedings of ICPhS 2019, Melbourne.

PHONETIC LESSONS FROM AUTOMATIC PHONEMIC TRANSCRIPTION: PRELIMINARY REFLECTIONS ON NA (SINO-TIBETAN) AND TSUUT'INA (ATHABASKAN) 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 cox.christopher@gmail.com 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 [19], 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 (Athabaskan). 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. Frederick Jelinek's quote "Anytime a linguist leaves the group the recognition rate goes up", dated c. 1988 [16, p. 83], is still alive in the oral tradition of the field of Natural Language Processing: the idea is that explicit (rule-based) modelling based on linguists' insights is outperformed by statistical modelling.

A lot is nonetheless at stake in collaborations between linguists and specialists of Natural Language Processing. Computer science research reveals that hand-crafted features can be meaningfully integrated in deep learning [24], with more promising results than under an 'end-to-end' black-box approach (see also [12]). This serves as a healthy reminder that "the goal of science is not wins, but knowledge" [21]. 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 (and more).

Specifically, phoneticians choose to focus on the forms of spoken language, suspending the ordinary flow of language, in which one *takes care of the sense, and the sounds take care of themselves* [5, p. 133]. Phoneticians deliberately stand back and dissect sounds with tools that allow for the separation of parameters which are not perceived separately by the human ear, such as fundamental frequency, spectral tilt, duration, intensity, and so on. So it does not appear necessary to labour the point that phoneticians have much to gain from reflecting on the workings of state-of-the-art speech processing tools: this can yield fresh insights and allow for new methods in phonetic research (see, for example, [22]). This point is illustrated in the present paper by reporting on lessons learnt when using an automatic phoneme transcription tool, `Persephone` (/**pər'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. In the perspective adopted here, emphasis is not placed on the tool's practical usefulness for language documentation [7], but on opportunities that it offers for phonetic research.

2. METHOD

2.1. About the phonemic transcription tool

In the interest of space, no attempt will be made here to describe the workings of the phonemic transcription tool used in this study, which implements a model similar to that of [10]. 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, which is referred to technically as "cross-validation", was applied to each of the texts in turn.

2.3. Choice of qualitative analysis

In this study, we conduct 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 (SINO-TIBETAN): THE ACOUSTIC SPECIFICITY OF LONG WORDS IN A PHONOLOGICALLY MONOSYLLABIC LANGUAGE

Yongning Na is a Sino-Tibetan language of Southwest China. A GitHub repository dedicated to Yongning 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 named `2018_08_StoryFoldCross-Validation`. 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 corresponding audio files are available from the Pangloss Collection [17].

3.1. Sample observations: an unusually high error rate on a quadrisyllabic proper name

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).

ɻ̍˩ tʂʰe˧ ɖɯ˩ mɑ˩ ɻ̍˩ tʂʰe˧ ɖɯ˩ mɑ˩ pi˧ dʑo˩
æ̃˩ tʂʰe˧ ɖɯ˧ mæ˩ tʂʰɯ˧ bi˧ mæ˩ pi˧ dʑo˩

(1) **ɻ̍˩tʂʰe˧ ɖɯ˩mɑ˩ pi˧ -dʑo˩**
Erchei_Ddeema to_say TOP
"'Erchei Ddeema! Erchei Ddeema!' [she] called out." (Sentence 13 of the narrative *BuriedAlive2*)

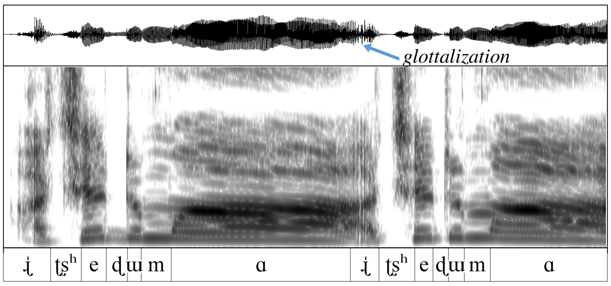
In this short example, quite a few transcription errors occur, all of them on the two occurrences of the proper name 'Erchei Ddeema' (the name of one of the characters in the story), phonemically /**ɻ̍˩tʂʰe˧ ɖɯ˩mɑ˩**/. In view of the tool's overall low phoneme error rate (on the order of 17%), it is striking to find nine errors on tones and consonants in a sequence of just eight syllables. To follow up on this initial observation, all occurrences of this name in the text were examined, and it turned out that none was devoid of mistakes: see Table 1. Phonemes are separated by spaces for visual clarity.

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

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

The first syllable, a syllabic approximant /**ɻ̍**/, is identified as a vowel in six cases, i.e. there is not enough acoustic evidence of retroflexion for identification as /**ɻ̍**/. (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 (at second occurrence of the name in sentence 14, and in sentence 77) 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 [14, p. 3218]. This glottalization may be in part responsible for lack of detection of the [**ɻ̩**] despite the presence of (admittedly slight) hints such as a final decrease in the third formant.

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



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 the syllable as [**ʈʂʰʐ̩**] (phonemically /**ʈʂʰɯ**/). 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 (˩). This reflects an acoustic fact: 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 [11]. These roots combine anew into disyllables through compounding and affixation, so that disyllables are widely attested, and combine, in turn, into longer words. Words of four syllables or more make up about 6% of a 3,000-word lexicon [18] and their frequency of occurrence in the 27 texts is similar (5.5%). Quadrisyllables are thus marginal in the data. This fact is held to be key to the errors shown in Table 1: the acoustic model tends to 'overfit' to the statistically more common type (monosyllabic or disyllabic morphemes, with limited phonological material, and consequently articulated with precision), to the detriment of the less common type (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 an interesting point is that 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. In future, we plan to train acoustic models with input data that contain word boundaries. This will allow for quantitative comparison of error rates when word boundaries are taken into account. 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.

4. TSUUT’INA (ATHABASKAN): REVEALING THE PHONEMIC VALIDITY OF AN ORTHOGRAPHIC CONTRAST

Tsuut’ina, an Athabaskan language, has four vowels, *i, a, o, u* (IPA: /**ɪ a ɒ ʊ**/) [15, 6]. A recent acoustic study based on the speech of one consultant concludes that these four vowels are still distinct phonemes, even though the acoustic distance between /**a**/ and /**ɒ**/ is small [2]. When recording the materials used in the present study, the linguist's impression was that the consultant was not producing this contrast in a consistent, recognizable way in spontaneous speech. (It should be noted that the language is highly endangered and not a few of the remaining speakers have limited fluency.) But 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 representation, not a string of IPA symbols. (The orthography is phonemic in orientation, and does not have considerable time depth.) The existence of all the orthographic contrasts in the consultant's speech was not verified systematically.

Interestingly, the acoustic model does a surprisingly good job of distinguishing the two hypothesized phonemes, /**a**/ and /**ɒ**/. This offers evidence that the consultant still makes the distinction: were the two phonemes merged in his speech, the statistical model would not be able to distinguish them consistently.

Use of automatic phonemic transcription thus has the side benefit of offering evidence on a difficult aspect of the Tsuut'ina phonemic system. Before this piece of evidence came to light, the transcribing linguist had been considering leaving aside the /**a**/-/**ɒ**/ vowel contrast from new transcriptions, being unsure whether it was still extant.

5. FUTURE WORK

The preliminary reflections set out here are so simple that they may seem unimpressive. But we now plan to refine them through innovative methods.

5.1. Attempting to extract knowledge from the acoustic model

A perspective for future research consists in 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 bring to light new knowledge about acoustic phonetics and phonology. This 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 [23].

In addition to vowels and consonants, Persephone also transcribes tones and prosodic boundaries, as provided in the training corpus. Prosodic boundaries are known to be cued by fundamental frequency, duration, phonation type, and fine detail in the articulation of vowels and consonants [8]; it would be interesting to find out which cues have greatest weight in identification of these boundaries by the transcription tool.

Due to the nature of the statistical models, known as 'artificial neural-network models,' it is not easy to look under the hood and 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 [20]. The use of such methods has the potential to amplify the insights the tool of speech recognition technology can provide to the phonetic sciences.

5.2. Connecting language processing with other branches of the phonetic sciences

Among the phonetic sciences, speech processing is most strongly linked to acoustic phonetics, because computer algorithms typically take an acoustic signal as input. Acoustic phonetics has strong claims to a key role among the phonetic sciences, but an audio-only perspective would clearly be an oversimplification. Focus on acoustics should not lead to overlooking the multimodal nature of human communication [13] or the fundamental importance of understanding the *biological* grounding of speech (e.g. [9]). Thus, to bridge the divide between Natural Language Processing, linguistic fieldwork (language documentation and conservation), and the phonetic sciences, new tools such as Persephone need to be integrated into a broader environment, reflecting the diversity of perspectives on language and speech.

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 and friends. 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] Carroll, L. 1866. *Alice's adventures in Wonderland.* Appleton.

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

[7] 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.

[8] Georgeton, L., Fougeron, C. 2014. Domain-initial strengthening on French vowels and phonological contrasts: Evidence from lip articulation and spectral variation. *Journal of Phonetics* (44), 83–95.

[9] Gick, B., Allen, B., Roewer-Després, F., Stavness, I. 2017. Speaking tongues are actively braced. *Journal of Speech, Language, and Hearing Research* 60(3), 494–506.

[10] 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.

[11] 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.

[12] 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.

[13] Keough, M., Derrick, D., Gick, B. 2019. Cross-modal effects in speech perception. *Annual Review of Linguistics* 5, 10.1–10.19.

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

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

[16] Martin, J. H., Jurafsky, D. 2009. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson/Prentice Hall.

[17] 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.

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

[19] 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.

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

[21] Sculley, D., Snoek, J., Wiltschko, A., Rahimi, A. 2018. Winner's curse? On pace, progress, and empirical rigor. *Proceedings of ICLR 2018, the Seventh International Conference on Learning Representations*.

[22] Shi, T., Kasahara, S., Pongkittiphan, T., Minematsu, N., Saito, D., Hirose, K. 2015. A measure of phonetic similarity to quantify pronunciation variation by using ASR technology. *Proceedings of the 18th International Congress of Phonetic Sciences.*

[23] 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.

[24] 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. *Proceedings of EMNLP 2018: 2018 Conference on Empirical Methods in Natural Language Processing.*