Towards the automatic processing of Yongning Na (Sino-Tibetan): developing a 'light' acoustic model of the target language and testing 'heavyweight' models from five national languages

Thi-Ngoc-Diep Do, Alexis Michaud, Eric Castelli

Proceedings of the 4th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU 2014), St Petersburg, May 16-18 2014, pp. 153-160

TOWARDS THE AUTOMATIC PROCESSING OF YONGNING NA (SINO-TIBETAN): DEVELOPING A 'LIGHT' ACOUSTIC MODEL OF THE TARGET LANGUAGE AND TESTING 'HEAVYWEIGHT' MODELS FROM FIVE NATIONAL LANGUAGES

*DO Thi-Ngoc-Diep** *MICHAUD Alexis**, ** *CASTELLI Eric**

*International Research Institute MICA, HUST – CNRS/UMI-2954 – Grenoble INP, Hanoi University of Science and Technology
**Langues et Civilisations à Tradition Orale (LACITO), UMR-7107 – CNRS, Paris 3-Sorbonne Nouvelle, Paris 4-Sorbonne, INALCO

ABSTRACT

Automatic speech processing technologies hold great potential to facilitate the urgent task of documenting the world's languages. The present research aims to explore the application of speech recognition tools to a little-documented language, with a view to facilitating processes of annotation, transcription and linguistic analysis. The target language is Yongning Na (a.k.a. Mosuo), an unwritten Sino-Tibetan language with less than 50,000 speakers. An acoustic model of Na was built using CMU Sphinx. In addition to this 'light' model, trained on a small data set (only 4 hours of speech from 1 speaker), 'heavyweight' models from five national languages (English, French, Chinese, Vietnamese and Khmer) were also applied to the same data. Preliminary results are reported, and perspectives for the long road ahead are outlined.

Index Terms— Acoustic models, automatic speech recognition (ASR), multilingual modelling, under-resourced languages, endangered languages, Yongning Na, Naish languages, language portability, statistical language modeling, crosslingual acoustic modelling and adaptation

1. INTRODUCTION

The present exploratory research is part of a long-term endeavour to tap the potential of automatic speech processing technologies to facilitate the urgent task of documenting the world's languages. This paper is aimed at the interdisciplinary readership of the SLTU (Spoken Language Technologies for Under-resourced Languages) workshop series. Linguists who would like to skim the paper for information about what linguistic findings can be made in the early stages of speech-recognition software development are advised to read §1.2-1-3, then jump straight to section 4 ("Linguistic findings"). Speech processing specialists looking for the key facts about what has been done so far with the Yongning Na data can start from §2.3 (presentation of the online data set), and continue into §3 ("Method").

As an introduction, some reflections will be set out about current bottlenecks in language documentation and the contribution that speech recognition tools could make.

1.1. Bottlenecks in language documentation: transcription and alignment

The necessity to document the world's languages is now well known to linguists and the general public. In recent years, a number of archives have been created to address this major need of the linguistic community – including the Academia Sinica Collections, AILLA (Univ. of Texas), ELAR (SOAS, London), the Language Archive at the Max Planck Institute for Psycholinguistics, and PARADISEC (ANU/Univ. Sydney). The Open Language Archives Community (OLAC) and the Language Archive at the Max Planck Institute for Psycholinguistics list many other repositories. Each of these archives makes a contribution to the world-scale effort of documenting the diversity of spoken languages. However, while it becomes increasingly easy to collect large amounts of data, data annotation remains highly time-consuming. After a trained linguist has worked out the phonological system of a given language or dialect, and trained their ear to be able to transcribe with accuracy, they still need about one hour of work – usually in collaboration with a native language consultant – to transcribe every minute of recording. This "pernicious transcription bottleneck" [1] puts severe limitations on the size of the data sets available for languages without a writing system: languages for which the only available resources are those created by professional linguists. In documentation projects, the proportion of transcribed materials is often significantly less than one fourth of the total recorded audio data. Further enrichment of the

annotation, such as the addition of phoneme-level alignment (which would be extremely useful for phonetic studies), is seldom carried out, again because it would require more time than the busy schedule of fieldwork and linguistic analysis generally allows.

Automatic speech processing technologies could facilitate the realization of these tasks. The following subsections deal with phonemic alignment and document transcription, respectively.

1.2. A perspective for the short to mid term: forced alignment of manual transcriptions

Phoneme-level alignment of transcriptions can now be conducted in automatic mode for a range of English dialects. This allows for applications in research such as the FAVE (Forced Alignment and Vowel Extraction) Program Suite, which automatically aligns and extracts large quantities of vowel formant measurements from orthographically transcribed data [2]. Such automated measurements hold great promise for shedding light on disputed issues of phonology, such as vowel deletion in French [3]: a recent study relies on 4,000 tokens of words produced as variants with and without *schwa* in a corpus of radio-broadcasted speech [4]; the scale of the corpus allows for new empirical insights into this classical issue. The extension of such fine-grained investigation to all the languages for which transcribed recordings are available would lead to progress (i) in the study of language-specific issues of phonological analysis and (ii) in the typology and modelling of sound systems. Forced-alignement tools, taking as their input the manual transcriptions created by linguists, would be a major step in this direction. Given the current state of the technology, this can confidently be planned in the short to mid term.

1.3. A perspective for the long term: automatic speech recognition as a tool to facilitate transcription

Another desirable tool for the linguist would be a full-fledged speech recognition system for the target language, which would provide a draft transcription – and, ultimately, translations into other languages, using automatic translation tools [5] – whose manual correction would be less time-consuming than fully manual transcription. This perspective currently appears distant, however, given the number of obstacles to overcome. Speech recognition technology requires large amounts of transcribed data in the first place; available data sets are typically too small to allow for the training of software for speech recognition and synthesis. For instance, in the course of two decades, the Pangloss Collection, an open archive of endangered-language data [6], has reached the size of about 190 hours of recordings in 70 languages; about 1/3rd of the total (60 hours) have a full transcription and annotation. This makes the Pangloss Collection one of the world's largest archives of "rare" or "endangered" languages. But this size – on average: one hour per language – remains extremely small in comparison with the amount of data currently required to train systems of recognition and synthesis. As a result, Human Language Technologies remain restricted to a small fraction of the world's languages.

In view of the situation summarized above, a suggestion that has been made in the literature is to improve "the portability of speech and language technologies for multilingual applications, especially for under-resourced languages" [7], see also [8]. The present paper reports on preliminary work intended as a contribution to this research strand, focusing on automatic speech recognition (ASR) for an unwritten Sino-Tibetan language of Southwest China that has less than 50,000 speakers.

2. PRESENTATION OF THE TRAINING CORPUS

Before presenting the data set used as a training corpus for the present study, it may be useful to present some general observations about data collection in fieldwork.

2.1. How the data set develops: the choice of recorded materials is guided by the research topics

In fieldwork, *word lists* are elicited first, to work out the phonemic system. But *narratives* (such as folk tales, life stories, and explanations about traditional techniques) and *dialogues* are the backbone of linguistic documentation. They offer examples of the use of words in context, and constitute a reliable basis for an open range of research purposes.

In addition to these basic types of documents, recordings are guided by the issues encountered in research, and reflect the diversity of linguists' interests. In the case of the documents presented here, the balance is tilted towards phonetic/phonological topics, which the researcher had the interest to pursue in greatest detail. Specifically, the tone system of the target language of the present study – Yongning Na – is an area which calls for in-depth description: the language has a complex system of morphologically conditioned tone change [9]. Numerous elicitation sessions have therefore been devoted to an investigation of the tone system, each focussing on a specific issue, such as tone changes that take place when an object is associated with a verb.

There is therefore no sharp distinction between a documentation agenda and a research agenda: all documents recorded in the course of research are relevant additions to the online collection, gradually enriching the record.

2.2. Advantages of fieldwork conditions for collecting abundant and reliable data

Data collection is an underestimated challenge, and perhaps a weak spot of some current linguistic studies. It is obvious that the empirical basis of one's research is of paramount importance for all later stages. On the other hand, the importance of good communication with language consultants is not always recognized. The consultants' perception of the investigator's intentions exerts considerable influence on their behaviour [10]–[12]. In this respect, the fieldworker's experience may be useful to the "laboratory worker". Documents collected in fieldwork compare favourably in many respects with those collected in the lab. In fieldwork, the investigator can gain familiarity with her or his consultants. (The second author of this paper stayed in the village of Yongning about two months a year from 2006 to 2009; from August 2011 to October 2012, he worked with his main consultant on a day-to-day basis.) This allows for the thoughtful design of materials to be recorded.

The transcription and annotation of the Yongning Na data was entirely manual, as is common for fieldwork materials. For some languages, adding complete glosses at word level can be done semi-automatically; however, semi-automatic treatment is painstaking in languages that contain numerous homophonous words, making it sometimes more appropriate to do all the glossing by hand rather than wade through lists of homophonous lexical entries.

Creating reliable, fine-grained transcriptions and annotations for documents in less-documented languages is a labour of love, into which investigators and their consultants put great amounts of time and effort. The quality of these hand-made annotations is usually excellent. This is a good start in life for these resources, which can later be further enriched and used for a variety of purposes.

Less positively, the way language documentation is conducted under fieldwork conditions entails some drawbacks for speech processing, as the data set is not tailored to match the needs of the speech recognition algorithms.

2.3. Current state of the online collection, used as training corpus for an acoustic model: a single-speaker data set

This section presents the audio data set available for the target language, Yongning Na, also known as Mosuo [13]–[15]. It updates an earlier presentation of these data [16].

In the Glottolog inventory of languages, the language code of Yongning Na is yong1270; in the Ethnologue inventory maintained by the Summer Institute of Linguistics, its code is NRU. Na is a tonal language whose syllable structure is essentially CV [13]. As most of the Yongning Na data were collected from one speaker, it was decided to create a speaker-specific language processing tool, not a speaker-independent tool.

One hundred narratives, making up a total of eleven hours, have been recorded by this speaker since 2006. Eight texts with complete transcription and Chinese and French translations are now available online (two of them with English translation as well), corresponding to 1.3 hour. Sixteen more (also available online) are transcribed and translated but without word-level glosses. Phonetic/phonological elicitation sessions with full Chinese, English and French annotations amount to more than two hours (over 40 documents). For narratives, the annotations are synchronized with the recordings at a level loosely referred to as the Sentence, with an average duration on the order of three seconds. Phonological materials are made up of phrases or short sentences, and the time-alignment is based on these units, which are less than three seconds in length.

The resources (recordings and annotations) are freely available for online browsing and download, under a CreativeCommons licence. The Pangloss Collection is hosted in a broader repository named CoCoON, via which the resources are referenced by various search engines including OLAC and OAIster. The metadata follow Dublin Core/Open Language Archives Community standards. The Yongning Na data are presented on the following page:

http://lacito.vjf.cnrs.fr/archivage/languages/Na_en.htm

A technical description of archiving and web hosting falls outside the scope of this paper; see [6]. Let us simply mention that long-term conservation is guaranteed through a partnership with a perennial archiving institution: CINES [17].

3. METHOD

"Porting Human Language Technologies (e.g. a speech recognition system) to an under-resourced language requires techniques that go far beyond the basic re-training of the models" [7], as exemplified by previous work [18]. The presentation of the method followed in the present preliminary work aims to emphasize the adjustments required by the processing of the target language.

3.1. Issues encountered when using spontaneous speech as a training corpus

As explained in section 2, one half of the training corpus is made up of narratives, and the other half of elicited phrases recorded to document the morpho-phonology of Yongning Na. This results in a degree of heterogeneity in the data, as elicited materials are pronounced with a slower speaking rate. In terms of the continuum between hypoarticulated and hyperarticulated speech [19], elicited materials tend to be hyperarticulated, whereas spontaneous speech contains

rapid alternations between hypoarticulation – typically, for backgrounded words – and hyperarticulation – for words that carry a greater informational and/or affective load. Spontaneous speech also contains dysfluencies and loanwords.

3.1.1. Dysfluencies

The narratives that constitute one half of the training corpus contain, like any set of continuous, non-read speech, some dysfluencies. Two types of filled pauses are distinguished in the narratives, those with the mouth open, transcribed as /əəə.../ (200 occurrences), and those with the mouth closed, transcribed as /mmm.../ (170 occurrences). One of the shortcomings of the preliminary version of the recognition system developed for Na is that it ignores filled pauses. In future versions, attempts will be made to identify various types of pauses, and indicate them in the automatic transcription.

Passages that were said by mistake (speech repairs) are transcribed in the annotation, aiming to approach as closely as possible (in consultation with the speaker) the actual flow of speech on the recording: for instance, the angle brackets in /ɖɯɬ-sɑʅ | <leɬ-ʂ...> ɖɯɬ-məɬ-ʁoɬ-tsɯȷ | -mɣȷ/ indicate that <leɬ-ʂ...> was said by mistake. In the first test reported here, the angle brackets were simply removed, and the dysfluent data were fed into the acoustic model along with canonical realizations. In later work, they will need to be treated separately, to avoid the presence of noncanonical realizations in the training data.

3.1.2. Loanwords

Minority languages in today's world are subject to increasing pressure from national languages and other socially dominating language varieties. This is reflected in the presence of loanwords. While older loanwords tend to be integrated into the language's phonological system, in cases of bilingualism loanwords tend to introduce new sounds and new combinations of sounds. This increases the inventory of sounds, and the number of choices among which the speech recognition algorithms have to make a choice. In the case of Na, recent borrowings from Mandarin Chinese introduce new sounds, such as /ẽ/ in /ɕiɬɕjẽɬ/~/ɕiɬɕɥẽɬ/ 'fresh' (Chinese 新鲜). A more frequent case is that Chinese borrowings introduce new combinations among phonemes: for instance, in /sjæɬpʰjɛɬ/ 'photograph' (Chinese 相片), the first syllable combines an initial /s/ with a rhyme /jæ/: this constitutes an unattested combination in Yongning Na. The second syllable contains a rhyme /jɛ/ which is altogether unattested in Na. In the initial version of the acoustic model, the sound /ɛ/ was simply merged with /e/; as the model is segment-based and not syllable-based, it was not necessary to face phonotactic issues, but these will need to be addressed at later stages.

3.2. Technical operations for building the 'light' acoustic model

3.2.1. Conversion of the recordings and transcriptions

The original recordings are WAV files, 24-bit, 44,100 Hz. One fifth consists of mono audio files, recorded with an AKG C-535EB or AKG C-900 microphone. Another fifth are stereo files with an audio channel and an electroglottographic channel. (Electroglottography is the ultimate reference for measuring fundamental frequency; it also allows for the evaluation of other glottal parameters [20]–[22].) The rest (60%) are stereo audio files comprising, in addition to the signal from a table microphone, a signal from a Sennheiser HSP 2 omnidirectional headworn microphone.

The electroglottographic recordings were not used so far; they will be useful for the analysis of tone, not undertaken here (see §5.1). The audio recordings were downgraded to 16-bit, 16,000 Hz to comply with the requirement on the input to CMU Sphinx. Stereo files were demultiplexed and fed into the training tool as if they constituted different data sets.

The annotation is logically structured text, in XML format. Narratives have the following structure: a TEXT is divided into S (sentences in a loose sense), which are divided into W (words) [23]. These data were converted to plain text, retaining only the sentence-level transcription and time codes.

3.2.2. Construction of an acoustic model, and speech recognition tests

An acoustic model was created by feeding the training data into CMU Sphinx. This 'light' model was trained on a small data set: only 4 hours of speech from 1 speaker. The resulting tool was then tested on a recording by the same speaker with the same linguistic and technical properties: a narrative recorded in a quiet room with the same equipment as the training corpus.

3.3. Application of 'heavyweight' recognition systems from other languages

In addition to the 'light' acoustic model developed for Yongning Na, 'heavyweight' recognition systems from five other languages (English, Mandarin Chinese, French, Vietnamese [24], [25], and Khmer [26]) were applied to the Na data, to determine to what extent Na sounds similar to sounds found in these five languages could be accurately recognized. A multilingual acoustic model for these five languages was built on the basis of [26]. This multilingual acoustic model combines five monolingual acoustic models. The total number of phones is 194: 40 English phones, 43 French phones, 34 Mandarin phones, 41 Vietnamese phones, and 36 Khmer phones.

The idea was to use the output from this 'heavyweight' multilingual model by applying rule-of-thumb acoustic correspondences between the target language's sounds and those of the five national languages selected. However, the error rate at first pass was extremely high. This is due in part to speaker characteristics: it was observed when doing spectrographic analysis of data from the Yongning Na speaker that her relatively high-pitched voiced, her speaking rate (often rapid) and the F-pattern of her small-sized vocal tract were a challenge to experimented spectrogram readers. In future work using speaker-independent recognition tools from other languages, it appears advisable to devise and apply speaker adaptation modules, following a strand of current research in automatic speech recognition [27], [28].

3.4. An example of the results

Figure 1 shows a sample of the acoustic data to which the recognition tool was applied.

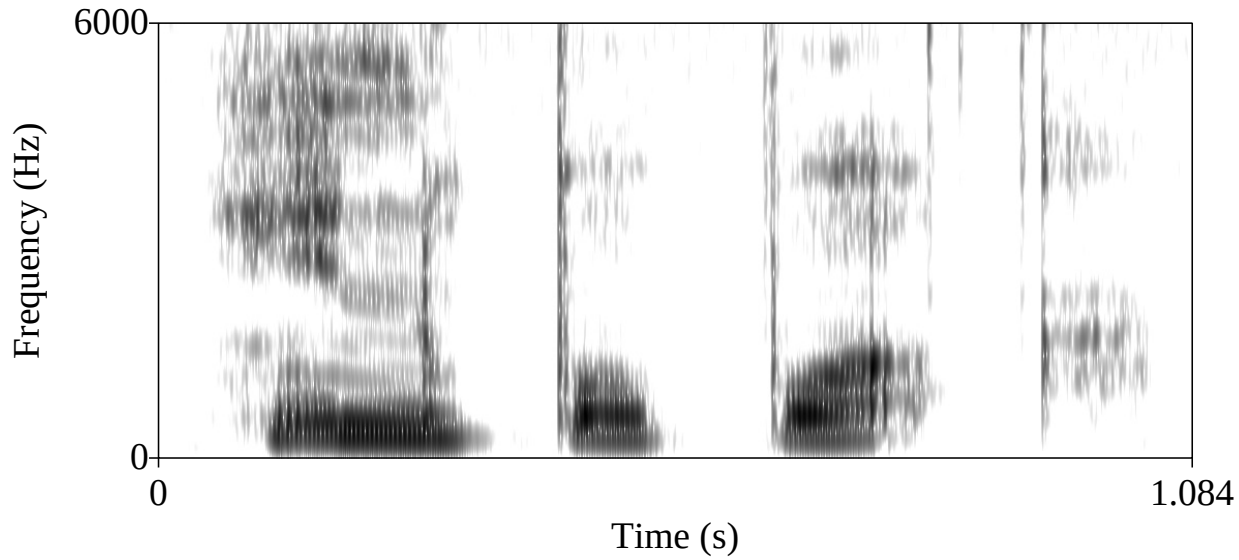

Fig. 1. A sample of Na data: /hĩ˧-ɳɯ˩ | pɤ˧~pɤ˧/

The passage shown in Figure 1 is /hĩ˧-ɳɯ˩ | pɤ˧~pɤ˧/, 'The people carried...'. It was recognized as /h ĩ ɳ ɯ p ɤ t o q v/ by the 'light' recognition tool developed for Na. Thus the first two syllables were correctly identified, whereas the two syllables /pɤ˧~pɤ˧/ were identified as /p ɤ/ and /t o/ respectively, i.e. with a mistake concerning the place of articulation of the initial stop. The initial of the second token of /pɤ˧/ in Figure 1 can hardly be confused visually with a /t/, since it does not present the F2 transition from about 1,800 Hz down to the target F2 value of the syllable's rhyme which would be a telltale sign of /t/. By comparison, the 'heavyweight' reconstruction models devised for other languages (see §3.3) yielded grossly inaccurate results for the unfamiliar sequence /hĩ.ɳɯ/ (identified as /k ɥ n/), and for the vowel /ɤ/ (identified as /o/ in the first case, and as /w/ in the second), but they identified the two labial stops accurately, suggesting that the mistaken identification proposed for /p/ by the Na acoustic model should be avoidable if a larger training corpus is used. (The passage shown in Figure 1 was identified as /k ɥ n p o p w/.)

Finally, the sound on the right hand-side of the spectrogram in Figure 1 is a nonlinguistic sound: the unvoiced release of a glottal stop, which could be approximated as [ʔ̥]. The Na acoustic model identified it as a syllable made up of an unvoiced uvular sound and a /ɣ/ rhyme; in order to avoid such mistaken identifications, it appears useful to provide more fine-grained notations for non-speech sounds than was done so far in the training corpus.

4. INITIAL LINGUISTIC FINDINGS

This section presents a small sample of linguistic findings based on the application of the very first version of the Yongning Na phoneme recognition tool. It is likely to be the most rewarding section of this paper for linguists; it is hoped that it will also be of interest to engineers working on speech processing, as an illustration of scientific side benefits of language technology.

4.1. Side benefits of data preparation

Data preparation offered a chance to check that all the data conformed to the phonological description. At data conversion from XML to plain text (§3.2.2), a list of segments was produced, and compared with the list of sounds provided for the language (by the second author). This comparison brought out a handful of inconsistencies in the notation, such as the use of /ẽ/ for an interjection appearing in some of the texts. This prompted a return to the data, which revealed that these were in fact instances of 'yes' (canonical transcription: /ĩ/) that had been transcribed before the second author identified the nature of this morpheme. Systematic examination of the passages at issue revealed that this /ẽ/ was in most cases a response to a comment or yes/no question on the part of someone in the audience, confirming the interpretation of this morpheme as a sign of approval.

4.2. Allomorphic variation of grammatical morphemes

Cross-linguistically, grammatical morphemes tend to be hypoarticulated, by comparison with lexical morphemes. This is brought out clearly by identification mistakes of the speech recognition tool.

4.2.1. Fricatives and affricates

The topic marker /-ʈʂʰɯ˩/ tends to be identified as /ʂ ɯ/, with a fricative instead of an affricate, and without aspiration (e.g. in sentence 1 of the narrative Housebuilding2). This may be revealing of its customary hypoarticulation (although the algorithm's mistake could also be due to other causes, such as problems in the learning process). Such phenomena argue for a different encoding of grammatical words vs. content words.

4.2.2. Lack of identification of nasals in grammatical morphemes

Weaker phonetic realization may lead to some syllables being entirely overlooked by the recognition system. Nasal-initial grammatical morphemes are a case in point: they tend to be overlooked by the recognition algorithm when they are followed by an item with an initial nasal. For instance, the negation /məɬ-/ was not identified in the passage /əɬtsoɬ-məɬ-ɲi˩/ (in sentence 1 of the narrative Housebuilding2), for which the algorithm's output was /ə ts o ɲ i/.

This stimulates further reflections on the linguist's part concerning the hypoarticulation of grammatical morphemes. The fact that the negation in Yongning Na undergoes a relatively high degree of vowel harmony with the verb that follows had led to its analysis as /məɬ-/, with a neutral vowel (schwa) which was defined phonologically as the neutralization of all vowel oppositions [29, p. 70]: a vowel that does not stand in opposition to the others. But the output of the recognition algorithm suggests that weakening also affects the initial. This leads to a questioning of the phonological analysis, in which the initial was considered identical to the /m/ found in other environments whereas a special treatment was set up for the rhyme. In this light, the second author's current thinking is that weakening should be considered a characteristic of the entire syllable, applying to a segmental structure that remains identical with that of lexical morphemes (content words). In turn, this raises the issue of the negation's phonemic vowel; phonetic evidence and historical-comparative evidence both point to an analysis as /mɤɬ-/. This reanalysis constitutes a step forward: arriving at a higher degree of generalization, while dispensing with a cumbersome theoretical device.

5. CONCLUSION: CHALLENGES AND PERSPECTIVES FOR FURTHER WORK

It is too early at present to launch into a performance evaluation for the system; instead, this section recapitulates challenges and perspectives for future work, emphasizing fundamental linguistic issues that will need to be integrated into the speech recognition software.

5.1. Handling linguistic tone

Tone was left out of the scope of this preliminary work, for the simple technical reason that it is not handled in CMU Sphinx. This is a major shortcoming, because tone has a high functional yield in Yongning Na. The common ancestor of Yongning Na and all other Sino-Tibetan languages was non-tonal, and replete with consonant clusters [30]; dramatic phonological erosion took place, giving birth to tone [31]–[33]; this erosion is especially advanced in Yongning Na [34]. For the Yongning Na speech recognition system to progress towards lexical identification, it will be necessary to handle tone recognition.

5.2. The need for tools that fit the structure of the target language: syllable-based recognition tools

Some contextual information is taken into account by the CMU-Sphinx software package: the preceding and following sound. In the case of Yongning Na, however, it appears necessary to go beyond these default settings and experiment with syllable-based recognition tools in future: a salient phonological characteristic of Yongning Na – as also of the closely related language Naxi [35] – is the tight association of initial and rhyme within syllables, which are essentially of CV structure. Coarticulation within Na syllables is so strong that it is sometimes an issue how to analyze a given syllable into an initial and rhyme. Given

this state of affairs, it definitely seems interesting to attempt to identify entire syllables, rather than successive sounds.

5.3. Identifying intonational phenomena

Intonational phenomena were entirely overlooked in the first-pass acoustic model. This is clearly a major shortcoming, and a source of errors in the automatic identification of vowels and consonants. Vowel lengthening is a case in point: there is no phonemic opposition of vowel length in Na; the last lexical item in a prosodic group tends to be greatly lengthened. In the first version of the software, lengthened syllables are often misinterpreted as a sequences of two syllables, the second with a continuant initial such as /l/. This is due to the absence of a parameter for syllable length in the first-pass version of the system. There is so much intonationally conditioned variation in syllable length in Yongning Na (and in the closely related languages Naxi [36] and Laze [37]) that this issue will definitely need to be carefully considered in future work.

An important task for our future work thus consists in finding ways to integrate intonational phenomena into the system. It is well-known that details in the articulation of lexically distinctive units (consonants, vowels, and stress or tone) carry intonational information, reflecting phrasing and prominence patterns [38]–[40] as well as speaker attitudes. In the long run, it would be desirable for the recognition system to identify intonation-conditioned phenomena: locating each recorded token within the acoustic space of variation of the corresponding phoneme, and deducing properties such as the degree of strength with which the token is articulated (about the "strength coefficient" as a relevant parameter for speech processing, see [41]).

There already exists some information about intonation in the documents used as training corpus, which contain:

- systematic indications of prosodic-group boundaries (by the symbol [|]), reflecting the division of the utterance into tone groups [42]
- indications about information structure: focus-marking, indicated by addition of the letter F after the word at issue; and emphatic stress, transcribed as an upward arrow ↑ before the syllable at issue [43]
- punctuation marks: commas, colons, semi-colons, full stops, exclamation marks, and interrogation marks.

These pieces of information were not used in the first-pass training of the acoustic model. They are all ultimately relevant for modelling, however, so that sources of variation across realizations of the same sound can be identified and analyzed, instead of treating variability as random.

6. ACKNOWLEDGMENTS

Many thanks to the Yongning Na language consultants and friends, and to Martine Toda and two anonymous reviewers for useful comments. We gratefully acknowledge financial support from CNRS through the PEPS (Projets Exploratoires Premier Soutien) grant *APRIL*, and from *Agence Nationale de la Recherche* through the projects *HimalCo*, ANR-12-CORP-0006, and *LabEx EFL*, ANR-10-LABX-0083 – Investissements d'Avenir.

7. REFERENCES

[1] S. Bird, "Androids in Amazonia: recording an endangered language (Blog entry)," 2013. [Online]. Available: http://theconversation.com/androids-in-amazonia-recording-an-endangered-language-12865. [Accessed: 21-Jan-2014].

[2] I. Rosenfelder, J. Fruehwald, K. Evanini, and J. Yuan, *FAVE (Forced Alignment and Vowel Extraction) Program Suite*. 2011.

[3] A. Martinet, "Nature phonologique d'e caduc," in *Papers on linguistics and phonetics in memory of Pierre Delattre*, The Hague: Mouton, 1972, pp. 373–379.

[4] A. Bürki, E. Mirjiam, C. Fougeron, C. Gendrot, and U. Frauenfelder, "What affects the presence versus absence of schwa and its duration : A corpus analysis of French connected speech," *Journal of the Acoustical Society of America*, vol. 130, no. 6, pp. 3980–3991, 2011.

[5] T. N. D. Do, E. Castelli, and L. Besacier, "Mining Parallel Data from Comparable Corpora via Triangulation," in *Proceedings of International Conference on Asian Language Processing - IALP 2011*, Penang, Malaysia, 2011.

[6] B. Michailovsky, M. Mazaudon, A. Michaud, S. Guillaume, A. François, and E. Adamou, "Documenting and researching endangered languages: the Pangloss Collection," *Language Documentation and Conservation*, 2014.

[7] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Communication*, vol. 56, pp. 85–100, 2014.

[8] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, no. 35, pp. 31–51, 2001.

[9] A. Michaud, "The tone patterns of numeral-plus-classifier phrases in Yongning Na: a synchronic description and analysis," in *Transhimalayan Linguistics*, N. Hill and T. Owen-Smith, Eds. Berlin: De Gruyter Mouton, 2013, pp. 275–311.

[10] C. Bowern, *Linguistic fieldwork: a practical guide*. Basingstoke [England] ; New York: Palgrave Macmillan, 2008.

[11] P. Newman and M. Ratliff, *Linguistic fieldwork*. Cambridge: Cambridge University Press, 2001.

[12] U. Mosel, "Field work and community language work," in *Essentials of language documentation*, J. Gippert, N. P. Himmelmann, and U. Mosel, Eds. Berlin/New York: de Gruyter, 2006, pp. 67–83.

[13] A. Michaud, “Phonemic and tonal analysis of Yongning Na,” *Cahiers de linguistique - Asie Orientale*, vol. 37, no. 2, pp. 159–196, 2008.

[14] L. Lidz, “A descriptive grammar of Yongning Na (Mosuo),” University of Texas, Department of linguistics, Austin, 2010.

[15] A. Michaud and Latami Dashi, “A description of endangered phonemic oppositions in Mosuo (Yongning Na),” in *Issues of language endangerment*, T. De Graaf, Xu Shixuan, and C. Brassett, Eds. Beijing: 知识产权出版社 (Intellectual property publishing house), 2011, pp. 55–71.

[16] A. Michaud, A. Hardie, S. Guillaume, and M. Toda, “Combining documentation and research: Ongoing work on an endangered language,” in *Proceedings of IALP 2012 (2012 International Conference on Asian Language Processing)*, Hanoi, Vietnam, 2012, pp. 169–172.

[17] “CINES: Centre Informatique National de l’Enseignement Supérieur. Overview -- Long-term data preservation -- Training Workshops -- Services -- HPC-Europa2.” [Online]. Available: http://www.cines.fr/. [Accessed: 15-Aug-2011].

[18] G. Caelen-Haumont, S. Sam, and E. Castelli, “Automatic labeling and phonetic assessment for an unknown Asian language,” presented at the International Conference on Asian language processing - IALP 2011, Penang, Malaysia, 2011, pp. 260–263.

[19] B. Lindblom, “Explaining phonetic variation: a sketch of the H&H theory,” in *Speech production and speech modelling*, W. J. Hardcastle and A. Marchal, Eds. Dordrecht: Kluwer, 1990, pp. 403–439.

[20] P. Fabre, “Un procédé électrique percutané d’inscription de l’accolement glottique au cours de la phonation: glottographie de haute fréquence,” *Bulletin de l’Académie Nationale de Médecine*, vol. 141, pp. 66–69, 1957.

[21] N. Henrich, C. d’ Alessandro, M. Castellengo, and B. Doval, “On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation,” *Journal of the Acoustical Society of America*, vol. 115, no. 3, pp. 1321–1332, 2004.

[22] R. J. Baken, “Electroglottography,” *Journal of Voice*, vol. 6, no. 2, pp. 98–110, 1992.

[23] M. Jacobson, B. Michailovsky, and J. B. Lowe, “Linguistic documents synchronizing sound and text,” *Speech Communication*, vol. 33 [special issue: "Speech Annotation and Corpus Tools"], pp. 79–96, 2001.

[24] V. B. Lê, “Reconnaissance automatique de la parole pour des langues peu dotées,” Ph.D., Université Joseph Fourier - Grenoble 1, Grenoble, 2006.

[25] H. Q. Nguyen, “Reconnaissance automatique de la parole continue à grand vocabulaire en vietnamien,” Ph.D., Université d’Avignon & Institut Polytechnique de Hanoi, 2008.

[26] S. Sam, “Vers une adaptation autonome des modèles acoustiques multilingues pour le traitement automatique de la parole,” Ph.D., Laboratoire d’Informatique de Grenoble & Institut Polytechnique de Hanoi, 2011.

[27] Z. Wang, T. Schultz, and A. Waibel, “Comparison of acoustic model adaptation techniques on non-native speech,” in *Proceedings of ICASSP 2003*, Hong Kong, 2003, pp. I:540–543.

[28] S. Sam, E. Castelli, and L. Besacier, “Online unsupervised multilingual acoustic model adaptation for nonnative ASR,” *ASEAN Engineering Journal*, vol. 1, no. 1, pp. 76–86, 2012.

[29] N. S. Trubetzkoy, *Grundzüge der Phonologie [Principles of Phonology]*. Prague: Travaux du cercle linguistique de Prague 7, 1939.

[30] M. Ferlus, “What were the four divisions of Middle Chinese?,” *Diachronica*, vol. 26, no. 2, pp. 184–213, 2009.

[31] A.-G. Haudricourt, “Comment reconstruire le chinois archaïque,” *Word*, vol. 10, no. 2–3, pp. 351–364, 1954.

[32] A.-G. Haudricourt, “Bipartition et tripartition des systèmes de tons dans quelques langues d’Extrême-Orient,” *Bulletin de la Société de Linguistique de Paris*, vol. 56, no. 1, pp. 163–80, 1961.

[33] J. A. Matisoff, “Tibeto-Burman Tonology in an Areal Context,” in *Proceedings of the symposium “Cross-linguistic studies of tonal phenomena: Tonogenesis, Japanese Accentology, and Other Topics,”* S. Kaji, Ed. Tokyo: Tokyo University of Foreign Studies, Institute for the Study of Languages and Cultures of Asia and Africa, 1999, pp. 3–31.

[34] G. Jacques and A. Michaud, “Approaching the historical phonology of three highly eroded Sino-Tibetan languages: Naxi, Na and Laze,” *Diachronica*, vol. 28, no. 4, pp. 468–498, 2011.

[35] A. Michaud, “Three extreme cases of neutralisation: nasality, retroflexion and lip-rounding in Naxi,” *Cahiers de linguistique - Asie Orientale*, vol. 35, no. 1, pp. 23–55, 2006.

[36] A. Michaud and He Xueguang, “Reassociated tones and coalescent syllables in Naxi (Tibeto-Burman),” *Journal of the International Phonetic Association*, vol. 37, no. 3, pp. 237–255, 2007.

[37] A. Michaud and G. Jacques, “The phonology of Laze: phonemic analysis, syllabic inventory, and a short word list,” *Yuyanxue Luncong 语言学论丛*, vol. 45, pp. 196–230, 2012.

[38] D. Erickson, “Effects of contrastive emphasis on jaw opening,” *Phonetica*, vol. 55, no. 3, pp. 147–169, 1998.

[39] D. Erickson, “Articulation of extreme formant patterns for emphasized vowels,” *Phonetica*, vol. 59, pp. 134–149, 2002.

[40] P. Keating, T. Cho, C. Fougeron, and C.-S. Hsu, “Domain-initial articulatory strengthening in four languages,” in *Phonetic Interpretation*, J. Local, R. Ogden, and R. Temple, Eds. Cambridge: Cambridge University Press, 2003, pp. 145–163.

[41] G. P. Kochanski and Shih Chilin, “Prosody Modelling with Soft Templates,” *Speech Communication*, vol. 39, no. 3–4, pp. 311–352, 2003.

[42] A. Michaud, “Phrasing, prominence, and morphotonology: How utterances are divided into tone groups in Yongning Na,” *Bulletin of Chinese Linguistics*, 2014.

[43] A. Michaud and M. Brunelle, “Information structure in Asia: Yongning Na (Sino-Tibetan) and Vietnamese (Austroasiatic),” in *Oxford Handbook of Information Structure*, C. Féry and S. Ishihara, Eds. Oxford: Oxford University Press, 2015.