Perceptual transcription and acoustic data: the example of /i/ in Yongning Na (Tibeto-Burman)

Alexis Michaud, Jacqueline Vaissière

Perceptual transcription and acoustic data: the example of /i/ in Yongning Na (Tibeto-Burman)[1]

Alexis Michaud & Jacqueline Vaissière***

*Langues et Civilisations à Tradition Orale (LACITO), CNRS/Univ. Paris 3
**Laboratoire de Phonétique et Phonologie (LPP), CNRS/Univ. Paris 3
alexis.michaud@vjf.cnrs.fr jacqueline.vaissiere@univ-paris3.fr

ABSTRACT

On the basis of fieldwork on a Tibeto-Burman language, Yongning Na, some reflections are offered on three interrelated issues: (i) To what extent does perceived allophonic variation correspond to articulatory/acoustic reality, to what extent does it merely reflect the investigator's perceptual expectations? (ii) How can acoustic data help select International Phonetic Alphabet symbols for the phonemic units brought out by distributional analysis? (iii) How can acoustic data help characterise the vowels and consonants encountered in fieldwork, both in a structural (language-internal) perspective, and in a cross-language perspective?

Key words: phonemics; acoustics; allophony; perception; Na/Naxi/Mosuo.

1. INTRODUCTION: USEFULNESS OF ACOUSTIC DESCRIPTION

The selection of International Phonetic Alphabet symbols for the phonemes of a language is essentially based on structural arguments and expert listening, rather than on acoustic analysis and modelling. This is true both for the most documented languages, because the descriptions that serve as standards are usually based on a tradition that predates acoustic phonetics, and for newly described languages, because acoustic analysis is seldom a priority in fieldwork. IPA notations, which are oriented towards the transcription of phonemic oppositions, offer some indications on articulatory characteristics, but they clearly do not encapsulate enough information to reproduce the sounds under study.

By contrast, the acoustic theory of speech production [2] allows for a representation of acoustico-perceptual characteristics of speech sounds by reference to the resonantial properties of the vocal tract: the F-pattern. Unlike IPA notation, a characterisation in terms of F-pattern is sufficiently detailed to serve as input for speech synthesis, via articulatory modelling [8]. (To be quite precise: the F-pattern needs to be supplemented by information on nasality, if any, and on phonation type.) In the belief that speech sounds can be characterised very accurately in terms of a target F-pattern and of the articulatory configuration used to attain this pattern, the second author of this article is currently developing a notation [11] to describe the phonemes of individual languages by reference to certain fixed acoustical

[1] Many thanks to Latami Dashi (拉他咪·达石) for his help and advice during fieldwork, and to the language consultants for their hospitality and their patience. Our thanks to Liberty Lidz for useful discussions about the phenomena that she observed in Luòshuǐ and elsewhere. The authors alone are responsible for any errors.

properties. This project is reminiscent of that of Daniel Jones in defining the cardinal vowels, insofar as the aim consists in providing a stable reference point for linguistic description. To our knowledge, the cardinal vowels are defined in articulatory terms, and the audio renderings proposed to illustrate them fluctuate somewhat, e.g. some of the cardinal vowels recorded by Peter Ladefoged are somewhat different from those of Daniel Jones [11]. It appears worth investigating in which ways an acoustic definition can supplement the IPA notations for a given language, and whether this facilitates cross-language comparisons.

As a tentative step towards implementing the research agenda outlined above, the present article exemplifies how acoustic observations can help overcome difficulties encountered in fieldwork. The object of study is the Na language (Tibeto-Burman family) as spoken in Yongning, China (Chinese coordinates: 云南省丽江市宁蒗彝族自治县永宁乡平静村). This language is also known as Mosuo (摩梭话) or Eastern Naxi (纳西语东部方言).

Preliminary remarks on the tonal system of Yongning Na

The tonal system of Yongning Na is currently under analysis. It appears to be based on two simple tones, H(igh) and L(ow), and two contours, L-to-M(id) and M-to-H. These four patterns, combined with a 'zero' value (absence of tonal specification), yield five possibilities for monosyllables, illustrated respectively by the words /ʐ̩wæ/ 'horse', /ŋ̩/ 'silver', /bu/ 'pig', /ɬ̩/ 'brains', and /lɑ/ 'tiger'. On disyllables, these five patterns are also attested; in addition, there also exist six more patterns, four of which arise through compounding (word-final L; word-final H; L+MH; L+H; LM+L) whereas the status of the last two remains unclear. A detailed description of this tonal system will be set out elsewhere. In the present article, tone is indicated in superscript before the word.

The acoustic analysis bears on the close, front vowel of Yongning Na, transcribed as /i/.

2. THE ISSUE: COARTICULATION, OR CATEGORICAL ALLOPHONY?

One of the issues encountered in fieldwork on Yongning Na was the notation of high vowels. The first, tentative notations included [e] as well as [i], [o] as well as [u]. Looking back at these transcriptions, it appears that close vowels [i] and [u] are in complementary distribution with close-mid vowels [e] and [o]. Table 1 shows their distribution.[1] Table 2 presents an inventory of rhymes.

Table 1. Distribution of close and mid-close vowels in the first field notations.

rhyme	[e], [o]	[i], [u]
initial consonant	retroflexes (all modes); dental fricatives and affricates	labials; dental stops; laterals; alveolo-palatals; velars; nasals; glottal; no initial

Table 2. A preliminary inventory of the rhymes of Yongning Na.

i		ɯ	u, ju
	v		
		ɤ, wɤ, jɤ	
	ɻ		
	æ, wæ, jæ		
		ɑ, wɑ	
+ /ĩ/, /ṽ/, /ũ/, /w̃ɤ/, /æ̃/, /ɑ̃/ after /h/			
+ /ʐ̩̃/ and /ũ/ as syllables			

The standard way of reporting on such observations is to describe the phoneme at issue as having two allophones, and to make a reasoned choice of one of these allophones to represent the phoneme: for instance, one could choose to transcribe /i/ for [i, e], and /u/ for [u, o]. Analysis as two allophones amounts to saying that the difference between the phonetic realisation of these phonemes in the two sets of consonantal contexts listed in table 1 goes beyond what can be predicted from

[1] Note that [7] does have a contrast between /i/ and /e/: L. Lidz analyses the 'fricative vowels' [ɻ] and [ʅ] as allophones of /i/ (e.g. [dzɻ], as in the verb 'to eat', is analysed as /dzi/, whereas we analyse it as /dzɯ/); the syllables where we have /dzi/, e.g. /dzi.lɯ/ 'wheat', are transcribed by [7] with an /e/.

coarticulatory properties of the preceding consonant: that the two contextual variants are phonetically different sounds.

This raises an important issue: does our classification of the realisations of one phoneme into two phonetic categories – e.g. transcribing [i] in [^{M}i] 'spot, pimple' vs. [e] in [^{H}ʈʂe] 'earth' – reflect an acoustic difference between these two realisations in Yongning Na, or does it merely reflect our perception as non-native speakers? Our perceptual expectations may detract from the precision and usefulness of the descriptions we produce.

Acoustic analysis can arguably provide evidence on this issue, supplementing auditory impressions. Untrained listeners go by the categories of their native languages (an effect of linguistic experience on the perception of front vowels was shown by [14]). It is not very likely that trained phoneticians can fully cast off this bias. Phonetic transcriptions done in the field are commonly considered to reflect a *phonetic* level, on top of which *phonemic* analyses can be built; however, field notations could be considered as *perceptual* materials, which call for a reflection on why the investigators, given their linguistic and scholarly background, chose a certain notation.

In the case of the present investigation, conducted by native speakers of French, the issue can be phrased as: What are the acoustic properties that led to the perception of the sounds [i] and [e]? To preview the result, it will appear

- that the phoneme at issue differs acoustically both from French /i/ and from French /e/;
- that our perception of [i] in some contexts and [e] in others can be explained by the coarticulatory effects of the preceding consonant.

3. OBSERVATIONS ON A FEMALE SPEAKER

3.1. Acoustic observations

The native language of the authors, 'Standard' ('Parisian') French, has phonemic /i/ and /e/, as part of a vocalic system which – in its more conservative varieties – has the full set of cardinal vowels, /i e ɛ a, u o ɔ ɑ/. In the notation currently being developed by the second author of this article, prototypical French /i/ corresponds to the acoustic configuration {palatal (⇑F3F4)^{3200Hz}}, i.e. F3 is maximal, it corresponds to a resonance of the front cavity, and since F3 comes in the vicinity of F4, the amplitude of one of the two formants, or of both, is reinforced. For a male voice this spectral prominence is located at about 3,200 Hz; however, the exact value in itself is less important than the concentration of energy – a phenomenon which the notion of F'2 (read "F2-prime") captures well [1]. French /i/ is "the most acute voiced, noise-free sound that a vocal tract can generate" [11]. This characterisation has recently been verified by the statistical analysis of formant values extracted from large databases: the average distance between F3 and F4 for /i/ is much smaller in French than in the seven other languages investigated by [5]. On the other hand, the acoustic configuration for French /e/ is one in which F1 is higher than the F1 of /i/, and where, for an adult male speaker, F2, F3 and F4 are roughly equidistant, i.e. not clustering together, unlike in /i/ (clustering of F3 and F4: ⇑F3F4) or /y/ (clustering of F2 and F3: ⇓F2F3).

An examination of spectrograms will offer insights into the Yongning Na sound originally transcribed as /i/. Spectrograms 1a, 2a and 3a show data from a female speaker aged about 55; 1b, 2b and 3b from a male speaker aged about 35. All items are pronounced in isolation.

The syllable /^{LM}i/ 'spot, pimple' begins with a semi-consonant (empty-onset filler). In Lijiang Naxi, the notation that has been proposed by [9] for syllabic /i/ is [ji]. Given the amount of friction noise observed, transcription of the onset as the fricative [ʝ] appears more adequate for Yongning Na.

The vocalic part of the syllable in spectrogram 1a is clearly diphthongised. The measurements in table 6 are taken at 1/4th of

the vowel, a time point chosen after [3:94], and at 5/6^{ths} of the vowel, where F1 is at its highest. Formant frequencies were estimated by using the PRAAT software package.

Table 6. Formant frequency measurements on spectrogram 1a. Time points are indicated relative to total vowel length.

time point	F1	F2	F3
1/4th	420	2740	3190
5/6^{ths}	500	2510	3310

The first formant rises in the course of the vowel, while the distance between F2 and F3 increases. During the first half of the vowel, much intensity is concentrated around 3000 Hz; this cluster appears to be made up of F2 and F3. A superficial examination would conclude that these characteristics are very much unlike those of French [i]: to a French ear, the perception of [i] requires, in terms of a male voice, an F3 frequency that is closer to that of F4 than to that of F2, or, said differently, a strong difference between F3 and F2 frequencies: F3-F2 ≈ 1000 Hz. On the other hand, from a perceptual point of view, it appears to be the frequency of the resulting spectral prominence that plays a key role in perception, whether such an acoustic prominence results from the clustering of F3 and F4 in a male voice or of F2 and F3 in a female voice. "It appears that in female and children's voices the relative role of individual formants may differ while an overall pattern aspect, yet to be defined, could be similar" [3:98]. (On the general issue of speaker normalisation in perception, and 'auditory *Gestalt* recognition', see [6].)

This is in keeping with our auditory impressions. When listening to the first and second halves of the vowel separately, our perceptual impression is [i] for the first half, [e] for the second half. It is known that a higher F1 can lead French subjects to perceive /e/ as opposed to /i/.

Spectrogram 2a shows a realisation of Yongning Na [^{M}tɕi] 'cloud' said in isolation by the same female speaker. Again, F1 rises (from 220 Hz to about 625 Hz) in the course of the rhyme, and F2 decreases (from 2420 to about 2170). F2 values are slightly lower than on spectrogram 1a; the amplitude of F2 movement is similar.

Spectrogram 3a shows the word for 'earth', initially transcribed as [^{H}ʈʂe]. (Note that its tone in citation form is the same as that of [^{M}tɕi] in spectrogram 2a: these two tonal patterns are neutralised in isolation.) Its vowel actually resembles acoustically those represented on spectrograms 1a and 2a. F2 is even higher than on spectrogram 2a. These facts are taken up in the following section.

3.2. Comments and perceptual hypotheses

The allophonic variation between [i] and [e] transcribed on the basis of auditory impressions does not match with the observations made on the female voice (spectrograms 1a, 2a and 3a): the vowel is realised by a fairly constant formant pattern. It may be described as an opening diphthong, somewhat like [i] at its beginning and like [e] at its end. The next step in analysis consists in finding out the reason for our perception of two distinct allophones.

Our perception of [e] in the word for 'earth', as in the other words with initial retroflex fricatives and affricates, may be due to characteristics of the consonant /ʈʂ/. The friction in /ʈʂ/ has an overall low frequency, with a peak at about 4,100 Hz as against 6,700 for /tɕ/ (spectrogram 2a). It is plausible that the perception of the initial portion of the vowel is influenced by the spectral characteristics of the initial consonant [for a review on perceptual context dependence, see 20:61]. In spectrograms 2a and 3a, the fricative noise continues until after voicing begins, i.e. into the beginning of the vowel; by itself, this is suggestive of a closed vowel – the area of greatest constriction is small, hence the production of turbulence noise by egressive airflow. However, since the retroflex (spectrogram 3a) has an overall lower spectrum than other fricatives – alveolo-palatal, as in 2a, or dental –, the superposition of its friction noise onto the beginning of the vowel may lower slightly the

perceived main spectral prominence of that vowel. Since the vowel at issue is acoustically in-between our reference points for [i] and [e], this small perceptual influence from the preceding consonant could account for the perception of [e] rather than [i]. This hypothesis was tested by means of a sketchy experiment, set out in section 4.

4. A PILOT CROSS-SPLICING TEST

Cross-splicing, made easy by digital sound processing, can shed light on issues such as the one set out in section 3. The initial /tɕ/ of spectrogram 2a was placed before the vowel in 3a. Our perceptual categorisation of the resulting stimulus confirms the expectations: after /tɕ/, we perceive this vowel as [i]. (This stimulus is available, together with the original sounds, at the internet address indicated in section 6.) Conversely, when the initial /ʈʂ/ of spectrogram 3a is placed before the vowel in 2a, this vowel is not perceived as [i] anymore, but as a more open front vowel, [e] or [ɛ].

This result can be extended to the [u]-vs.-[o] allophonic pairs of our initial notations.

5. CONCLUSIONS AND PERSPECTIVES

5.1. On the hypothesised allophonic variation of close vowels

The acoustic observations on the sounds initially transcribed as distinct allophones, [e] and [i], differ for the female speaker and for the male speaker.

In the data from the female speaker, a member of the older generation who speaks Yongning Na on a day-to-day basis, the contextual variants of the phoneme at issue are much closer acoustically than the first fieldwork notations suggested. This phoneme is an opening diphthong; to a foreign ear expecting a stable vowel quality, it can be perceived as [i] due to the high frequency of the spectral prominence found at vowel onset, or as [e] when the spectral properties of the preceding consonant have the effect of lowering the perceived spectral prominence.

On the other hand, in the data from the male speaker (spectrograms 1b, 2b, 3b), the acoustic realisation of this phoneme after a retroflex initial differs from its realisation in isolation and after a palato-alveolar fricative. The vowels in spectrograms 1b and 2b have a similar F-pattern, whereas 3b shows a rising F2. These data suggest that in the case of the male speaker, who is one generation younger and has much greater proficiency in Chinese, there is actually some allophonic variation of the close front vowel phoneme under investigation. Needless to say, this observation will need to be confirmed by examining more data, and studying more speakers of the same age.

5.2. Implications for fieldwork: usefulness of acoustic observations

It is well known that field workers need to be careful to describe the phonetic categories of the language they investigate in their own terms, rather than basing their judgments on analogies with sounds found in other languages that they know, such as their mother tongue or other dialects of the language under analysis. In the case of Yongning Na, the first author's earlier experience of fieldwork in 2002 and 2004 on a closely related language variety, Lijiang Naxi, probably exercised an unwanted influence: in Lijiang Naxi, there is a contrast between /i/ and /e/ (the latter actually closer to [ɛ]) – e.g. /mbeɬ/ 'village', /mbiɬ/ 'urine' –, and between /o/ and /u/ – e.g. /kʰoɬ/ 'noise, sound', /kʰuɬ/ 'door'. This may have unwittingly led to an expectation that Yongning Na, too, had closed vowel phonemes and mid-closed vowel phonemes.

The results in section 3 illustrate the fact that a simple acoustic analysis can offer some evidence to overcome the uncertainties and shortcomings of auditory impressions. On the other hand, these results also suggest that the acoustic study needs to be conducted in a systematic way, collecting data from several

speakers and building a corpus containing each phoneme in a variety of contexts, in order to obtain reliable results. This conclusion opens into a project which is outlined in section 5.3.

5.3. Issues of transcription: necessity of a systematic acoustic overview.

Concerning the notation of the high front diphthong studied in section 3, a decision can hardly be made in isolation, without considering the broader picture of the acoustic characteristics of the phonemes of Yongning Na. The choice between a notation as /ɪ/, /i/, or again a notation that would represent both the beginning and endpoint of the diphthong (such as /ɪᵉ/ or /iᵉ/) depends in part on how this sound relates to other vowel phonemes in the language, and in part on how its formant structure compares with the reference values available for an increasing number of languages: American English [10], Swedish [3:94], French, German [4]…

A consequence of the structural/functional notion of phonemic *system* is that the full picture of the language's contrasts should be taken into account before an adequate acoustic characterisation of individual sounds can be proposed. We therefore plan to produce a monograph on the acoustics of Yongning Na, on the basis of recordings by two women and two men.

The ultimate aim would be to propose a characterisation of each vocalic and consonantal phoneme by means of the fine-grained acoustic notation put forward by [11]. To this end, the more phonetic instruments can be taken into the field, the better: acoustic modelling should be combined with articulatory modelling, which requires more information than spectrograms can offer. A comparable acoustic output can be obtained by different articulatory configurations, so that hypotheses suggested by spectrographic analysis can only be verified by means of physiological data. We therefore plan to supplement audio data with multisensor data, such as oral and nasal airflow.

6. LINK TO SOUND FILES

The sounds corresponding to the spectrograms are available online from:
http://ed268.univ-paris3.fr/lpp/pages/EQUIPE/michaud/NAXI/Yongning_Na_HighFrontVowel.htm

7. REFERENCES

[1] Bladon, A.; Fant, G., 1978. A two-formant model and the cardinal vowels. *Speech Transmission Laboratory Quarterly Progress Status Report* 1, 1-8.

[2] Fant, G. 1960. *Acoustic theory of speech production*. 1960, The Hague/ Paris: Mouton.

[3] Fant, G., 1973. *Speech sounds and features*. Cambridge, Massachusetts: MIT Press.

[4] Gendrot, C.; Adda-Decker, M., 2005. Impact of duration on F1/F2 formant values of oral vowels. *Proceedings of Eurospeech/Interspeech 2005*, Lisboa, 2453-2456.

[5] Gendrot, C.; Adda-Decker, M.; Vaissière, J. 2008. Les voyelles /i/ et /y/ du français : aspects quantiques et variations formantiques. Proceedings of *Journées d'Etude de la Parole 2008*, Avignon, France.

[6] Johnson, K., 2004. Speaker normalization in speech perception. In *Handbook of Speech Perception*, D. B. Pisoni and R. E. Remez (eds). Oxford: Blackwell, 363-389.

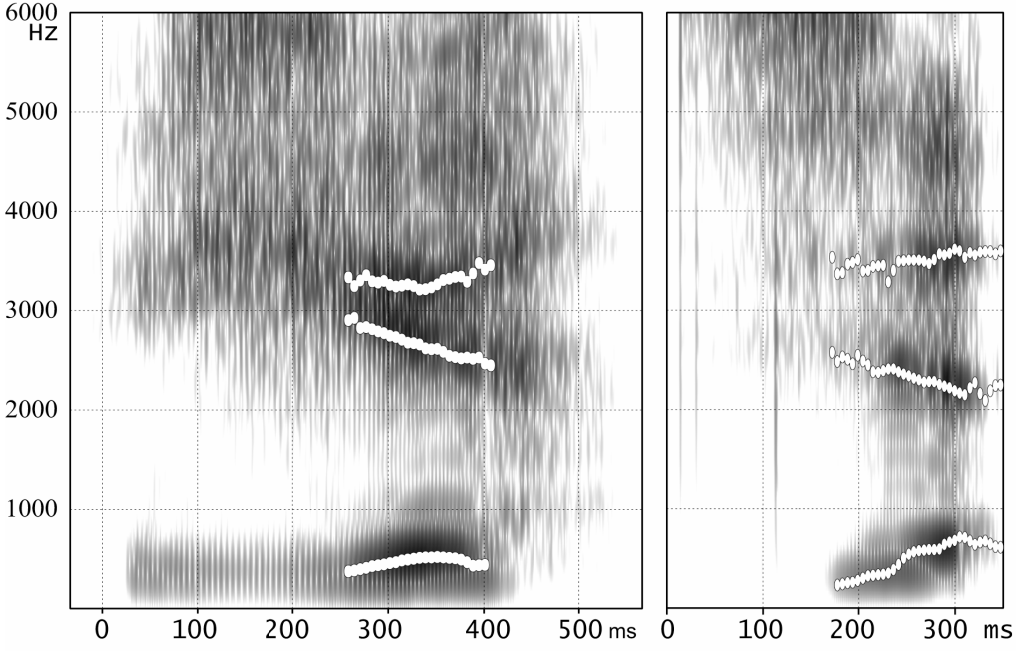
[7] Lidz, L., 2006. A synopsis of Yongning Na (Mosuo). Handout circulated at the 39th International Conference on Sino-Tibetan Languages and Linguistics, University of Washington, Seattle.

[8] Maeda, S., 1996. Phonemes as concatenable units: VCV synthesis using a vocal-tract synthesizer. In: *Sound patterns of connected speech. Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung 31*, A. Simpson and M. Patzod (eds.), Kiel, 127-232.

[9] Michailovsky, B.; Michaud, A., 2006. Syllabic inventory of a Western Naxi dialect. *Cahiers de linguistique - Asie Orientale* 35(1), 3-21.

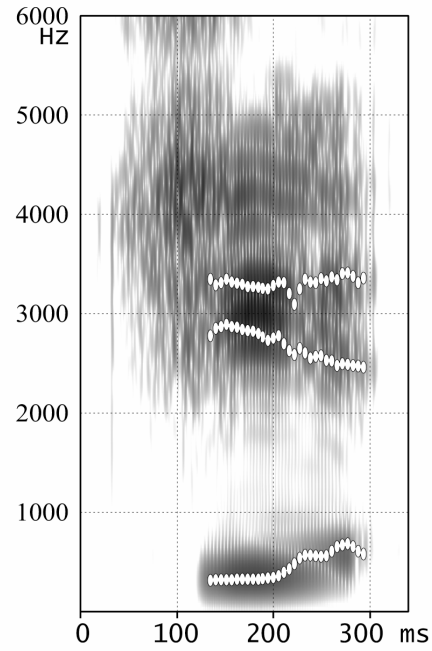
[10] Peterson, G.E.; Barney, H.L., 1952. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America* 24, 175-184.

[11] Vaissière, J., 2007. Area functions and articulatory modeling as a tool for investigating the articulatory, acoustic and perceptual properties of the contrast between the sounds in a language. In *Experimental Approaches to Phonology*, P.S. Beddor et al. (eds.). Oxford: Oxford University Press, 54-72.



Spectrogram 1a: /i/, realised as [ji]

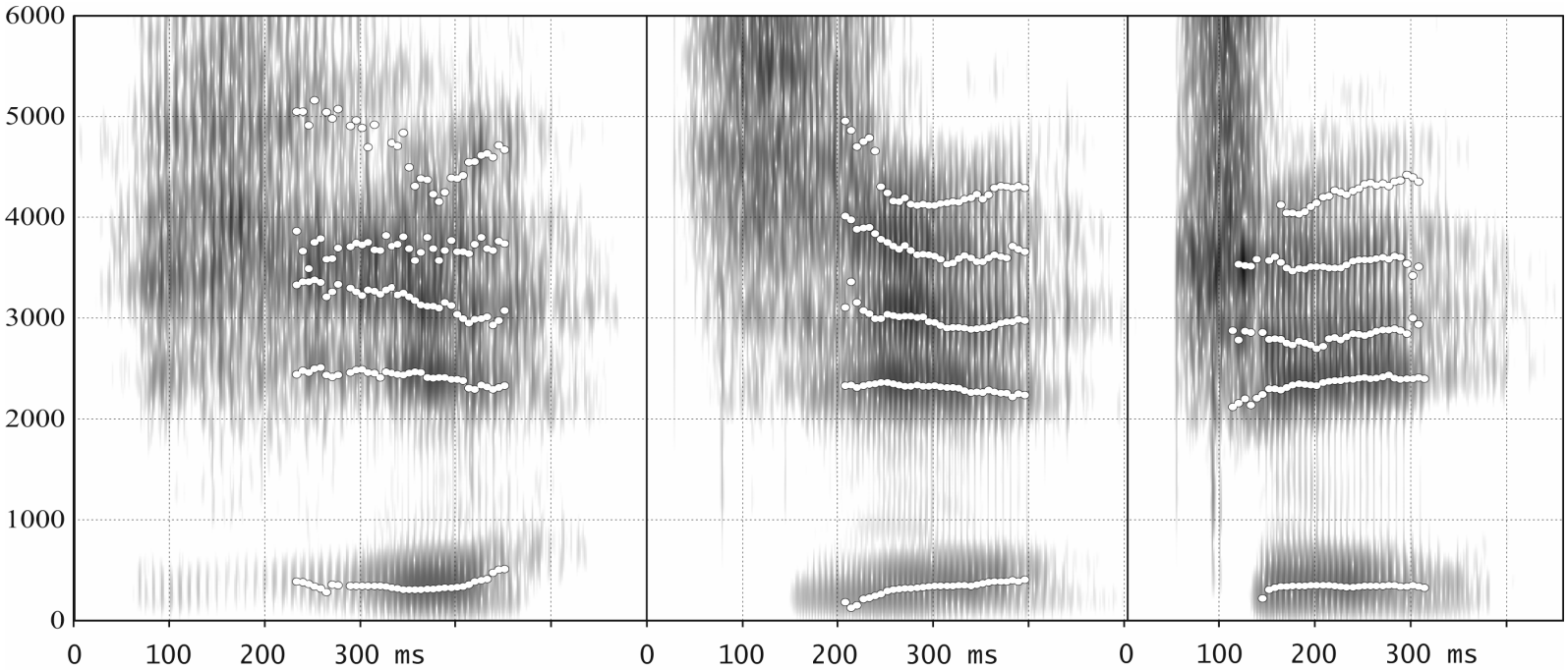
Spectrogram 2a: /tɕi/



Spectrogram 3a: /ʈʂi/ (initially transcribed [ʈʂɐ])

Spectrograms 1a, 2a and 3a
are based on data from an adult female subject.

The superimposed dots indicate the frequency of formants F1, F2 and F3 as evaluated by the software PRAAT.



Spectrogram 1b: /i/, realised as [ʝi] Spectrogram 2b: /tɕi/ Spectr. 3b: /ʈ͡ʂi/ (initially transcribed [ʈ͡ʂɐ])

Spectrograms 1b, 2b and 3b are based on data from an adult male subject.

The superimposed dots indicate the frequency of formants F1 to F5 as evaluated by the software PRAAT.