Spoken language identification by means of acoustic mid-level descriptors

Uwe D. Reichel[1], Andreas Triantafyllopoulos[1,2], Christopher Oates[1], Stephan Huber[1], Björn W. Schuller[1,2,3]

[1]*audEERING GmbH, Germany*
[2]*Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany*
[3]*GLAM – Group on Language, Audio & Music, Imperial College London, UK*
ureichel@audeering.com

Abstract: We introduce an acoustic mid-level feature (MLD) set derived from openSMILE low-level descriptors for the purpose of language characterisation and identification. The four languages targeted in this study are Georgian, Pashto, Kurmanji Kurdish, and Turkish. Language-dependent differences of these features will be discussed in terms of language typology. Furthermore, language identification by feed forward neural networks is comparatively evaluated for the MLDs and for openSMILE functionals, as well as for varying segment of analysis lengths. The best result 76.3% UAR was achieved for a joint feature set and for a minimum speech chunk length of 8 seconds.

1 Introduction

Spoken language identification (LID), the classification of a speaker's language has many applications in multi-lingual speech processing, such as multi-lingual speech recognition and dialog systems, in automatic machine translation, in audio information retrieval, or in forensics to detect a speaker's country of origin. Among the features used for LID are low-level acoustic signal representations such as mel frequency cepstral coefficients (MFCC) [1], high-level phonological representations such as phonotactics based on preceding phoneme recognition [2, 3], and prosodic features [4, 5]. Among commonly used classifiers for LID are Gaussian mixture models [6], support vector machines [7], and deep neural networks [8, 9].

The shortcoming of low-level acoustic features is that phonological concepts like the acoustic vowel space are decomposed into separate spectral features, the connection of which is to be re-discovered by the learner. The shortcoming of high-level phonological features is that their extraction is comparably laborious. The assessment of phonotactics for example requires the training of a phoneme recogniser. Furthermore, their extraction is vulnerable to errors made by the required models. Midlevel feature sets as proposed for speech analysis, e. g., by Ward [10] are likely to circumvent these shortcomings. Thus, in this study, we introduce a phonetically informed mid-level acoustic feature set that is derived from openSMILE low-level descriptors and that captures phonological and prosodic language aspects. Based on this set and on openSMILE functionals, feed forward networks are trained for the LID of four under-researched languages Georgian, Pashto, Kurmanji Kurdish, and Turkish. These languages stem from the three families of Kartvelian, Indo-Iranian, and Turkish languages, respectively. Selected language characteristics in terms of MLDs, and the LID performance in dependence of the feature set and segment length will be presented and discussed in the following sections.

2 Data

The data for the four target languages stems from four corpora produced within the IARPA Babel program for underserved languages: the IARPA Babel Georgian (kat) [11], Kurmanji Kurdish (kmr) [12], Pashto (pus) [13], and Turkish (tur) [14] language packs. The corpora contain between 190 and 214 hours of conversational and scripted telephone speech from a variety of environments (please see references for further details). All corpora are distributed by the Linguistic Data Consortium. The audio data is provided in 8 kHz 8 bit a-law encoding in Sphere format and in 48 kHz 24 bit PCM Wave format. All corpora contain a text-transcribed segmentation into speech chunks with a 1st to 99th percentile duration range from .48 to 13 seconds. All corpora are split into a training and a test partition with 281 249 training chunks and 40 548 test segments.

Before feature extraction, all audio files were converted to wav and uniformly downsampled to the lowest provided sample rate of 8 kHz. For each speech chunk, an acoustic feature vector was extracted. The text transcription was not used in the current study.

3 Feature extraction

3.1 openSMILE functionals

The first feature set used for the identification task is given by the openSMILE ComParE 2016 basic feature functionals, which was provided amongst others at the Interspeech 2016 Computational Paralinguistics challenge on native language detection [15]. It consists of various summary statistics of acoustic *low-level descriptors (LLDs)* derived by short-time analysis of the speech signal in 10 ms steps. The low-level descriptors comprise of spectral features, e.g. MFCCs, voice quality features like jitter and shimmer, as well as prosodic low-level features, namely F0 and energy. Summary statistics of these descriptors are calculated over each pre-segmented chunk in our corpora.

3.2 Mid-level descriptors

Based on the low-level descriptors introduced above, we derive a mid-level representation of the speech signal by (1) focusing on analysis windows on syllable nuclei only, and (2) aggregating features within the same vector and over time. These mid-level descriptors – a term already in use, e. g., by Ward [10] – allow for a more phonetically informed feature extraction. In this study, they refer to phonological and prosodic aspects of speech, namely vowel space, intonation, and rhythm.

Prosodic structure Both the phonological and prosodic dimensions are closely linked to vowels and syllables. Thus, MLDs were calculated only from LLDs in windows centered on syllable nuclei, which are generally vowels.

Within each predefined chunk, syllable nuclei were identified by a rule-based method described in [16] in terms of local energy maxima. Those local peaks that supersede a local and a global relative energy threshold, and that do not violate a minimum syllable duration constraint, are chosen as nuclei.

MFCC spread and flux Two key notions of the MLDs are *spread* and *flux*. *Spread* we define as the Euclidean distance of an LLD feature vector at time t in a speech chunk to its centroid which is calculated over the entire chunk. For MFCC vectors, this notion is closely related to the articulatory *enunciation* measure of [10], only that the latter is calculated over all voiced

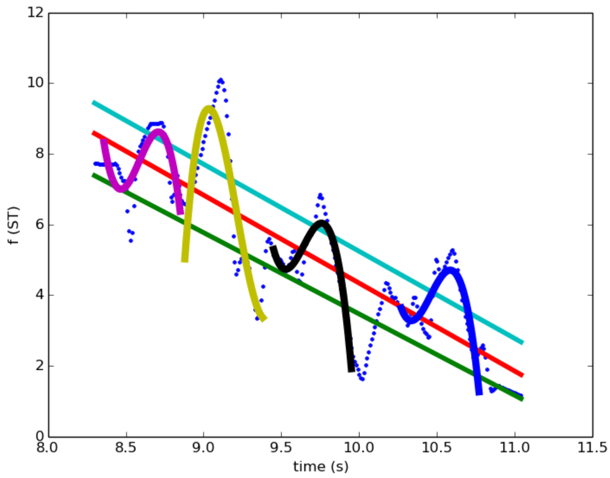

Figure 1 – Superpositional stylisation of F0 contours into register level (base-, mid-, and topline), register range (the distance between top- and baseline), and local F0 movements stylised by third-order polynomials.

frames, whereas we restrict the analyses to windows of length 30 ms around syllable nuclei to more precisely target the language-dependent vowel space size and articulatory precision [17].

Flux we define as the feature vector delta given by the Euclidean distance of two adjacent vectors. MFCC flux is intended to capture the notion of articulatory speed and might reflect the proportion of diphtongs in a language.

For both spread and flux within each speech chunk, the non-parametric summary statistics median, interquartile range, non-parametric versions of skewness and kurtosis, as well as the range between the 5th and 95h percentile were collected as features.

Intonation Intonation features are derived from a superpositional F0 stylisation consisting of a global register component and local F0 movements around syllable nuclei as proposed in [18] and illustrated in Figure 1. Smoothed F0 in semitones taken from the openSMILE feature table is first preprocessed by linear interpolation over outliers and voiceless segments. Then, within each speech chunk, the global F0 register level is calculated in terms of a base-, mid-, and topline as described in [18]. The F0 range was subsequently calculated as a regression line through the point-wise distances between top- and baseline. From these four lines the mean and slope were taken as MLDs which are intended to represent language-dependent F0 register characteristics.

After subtraction of the fitted midline the F0 residual contours around each nucleus were stylised within 200 ms windows by means of third order polynomials. The polynomial coefficients serve as a representation of a language's melodic patterns.

As for the MFCC vectors, we calculated spread and flux also for the polynomial vectors and added their summary statistics to the feature vector. Spread is intended to represent a language's F0 pattern variation, and flux is intended to represent the speed of its pattern changes.

Synchronisation of F0 and energy Within 200 ms windows around the syllable nuclei, we calculated the Spearman correlation and the RMS between the centered and scaled F0 and energy contours derived from openSMILE LLDs. This approach is similar to the F0-energy alignment proposed by [10]; however, here, the spread and flux calculation is restricted to the syllable nuclei. Since nuclei are defined in terms of energy maxima, a low correlation indicates a high variation of F0 peak positions relative to the syllable nucleus. Since this peak alignment was found to encode the information status of discourse referents, i.e. whether a referent is given or newly introduced in the discourse [19], this feature is intended to roughly reflect the degree to which a language makes use of intonation to mark information structure and focus.

Rhythm Rhythm is represented in terms of syllable rate (the number of extracted syllable nuclei per second), and in terms of the influence of the syllable oscillation on the F0 and energy contour. The latter is measured by decomposing the contours by means of a discrete cosine transform and by quantifying the relative weight of the cosine oscillations with frequencies around the measured syllable rate. The higher these weights, the higher the impact of the syllable oscillator (as opposed to the phonological foot oscillator) on the energy and F0 contour (see [20] for details). This weight is intended to represent the degree to which a language is syllable-timed rather than stress-timed [21].

Voice quality Finally, next to LLD integration into MLDs, non-parametric summary statistics were calculated from selected LLDs over windows of 30 ms length around syllable nuclei. The selected LLDs capture voice quality aspects as jitter, shimmer, harmonics-to-noise ratio, and zero crossing rate.

4 Language characteristics

4.1 Method

In a first exploratory data analysis, we identified features differing across languages by means of a Kruskal Wallis tests. 1164 out of the 1170 openSMILE and all 174 MLD features showed significant differences ($p < 0.01$). However, since this large amount of results was not corrected for type 1 errors, and since significant differences also emerge from large sample sizes, these results are rather to be taken as an initial rough overview that may lead to more targeted follow-up analyses on selected features. Statistics were calculated for speech chunks with a minimum duration of 8 s, which the results in section 5 suggest to be a good value for reliable feature extraction.

For the current study, we restrict the language characterisation and its linguistic interpretation to four exemplary MLDs: the median F0 shape spread, and the interquartile range of the polynomial coefficients in local F0 contour stylisation.

4.2 Intonation and focus marking

The median F0 shape spread and the interquartile ranges of the polynomial F0 contour stylisation coefficients represent the variation of local F0 contours around syllable nuclei. Figures 2 and 3 show a small, but consistent trend that F0 variation is lower for Georgian than for the other languages. Potential explanations for this trend are, first, that Georgian word stress is acoustically only very weakly marked (see [22] for a review). Second, in Georgian focus (the information centre of an utterance) is primarily marked by syntactic means: its position is left-adjacent to the verb [23]. Both the weak stress marking and the non-plasticity [24] in focus marking are likely to reduce the functional load of prosody so that less variation in intonation contours is needed.

5 Language prediction

5.1 Method

For training and testing the *training* and *dev* partitions of the IARPA-BABEL corpora were used, respectively.

We used feed forward neural networks with one hidden layer. The hidden layer had 100 nodes, and was followed by a LeakyReLU activation and dropout with probability 0.2. We trained separately for the openSMILE, the MLD, and the joint feature set, and varying minimum

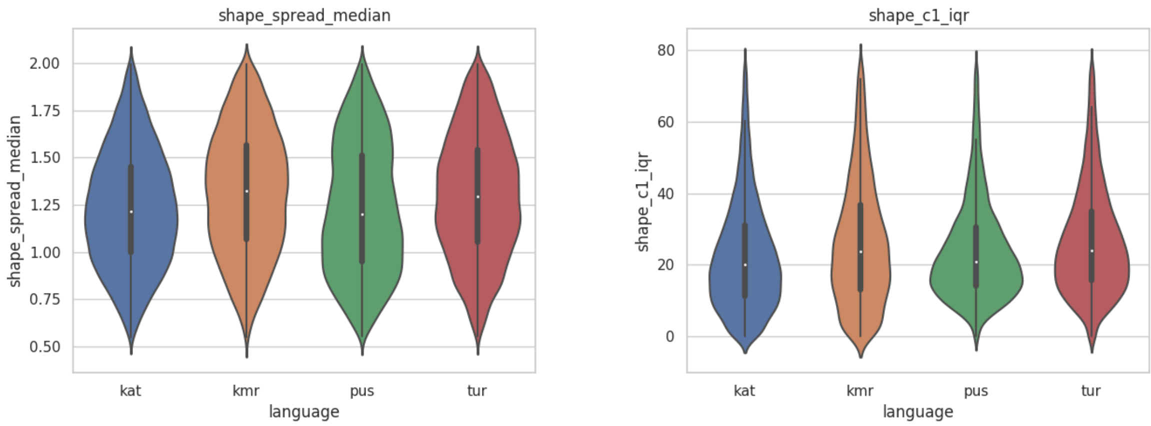

Figure 2 – **Left:** Median F0 shape spread for Georgian (kat), Kurmanji Kurdish (kmr), Pashto (pus), and Turkish (tur). **Right:** Interquartile range of the linear polynomial coefficient in local F0 contour stylisation. For both features, higher values indicate more varied F0 shapes in syllables. Distribution means differ significantly ($p < 0.01$).

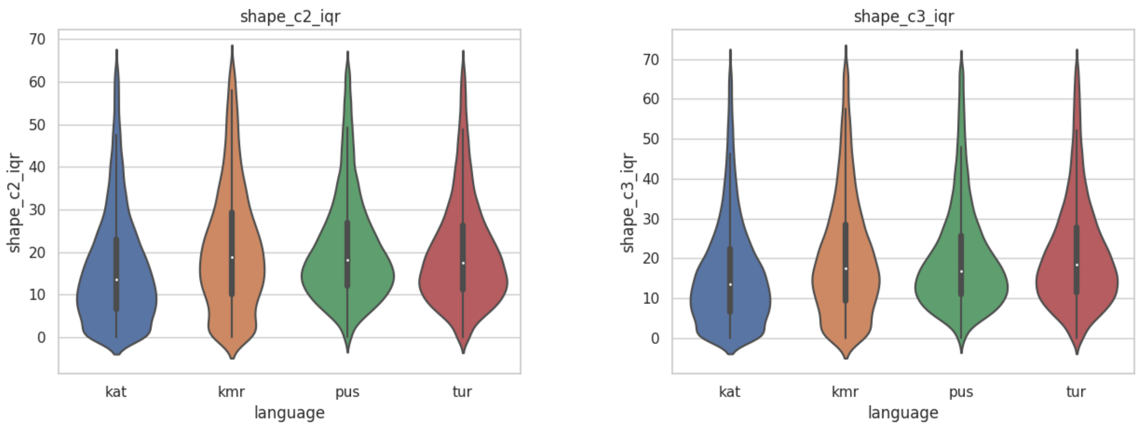

Figure 3 – Interquartile ranges of the quadratic (left) and hyperbolic polynomial coefficient (right) in local F0 contour stylisation. For both features, higher values indicate more varied F0 shapes in syllables. Distribution means differ significantly ($p < 0.01$).

durations of the speech chunks were taken into account. All networks were trained for a total of 40 epochs, starting with a learning rate of 0.01 and a batch size of 128. After 20 epochs, we decreased the learning rate to 0.001. Furthermore, we balanced the training data by randomly undersampling to the minority class.

5.2 Results

Figure 4 shows the unweighted average recall (UAR) values achieved on the held-out test data in dependence of the feature set and the minimum duration of the analysis chunks. A comparison of the feature sets in isolation yields an overall higher performance of the openSMILE functionals as opposed to the MLDs. Performance is highest for the joint set, for which a UAR up to 76.3 % is reached. For all sets, performance is highest at a minimum chunk length of 8 s. For this length, a detailed per-language and summary evaluation is presented in Table 1. Above 8 s, a performance decrease is observed mainly due to the decreasing amount of available training data.

Among the four languages the highest precision and recall is achieved for Georgian (kat) and Pashto (pus), as can be seen in Table 1.

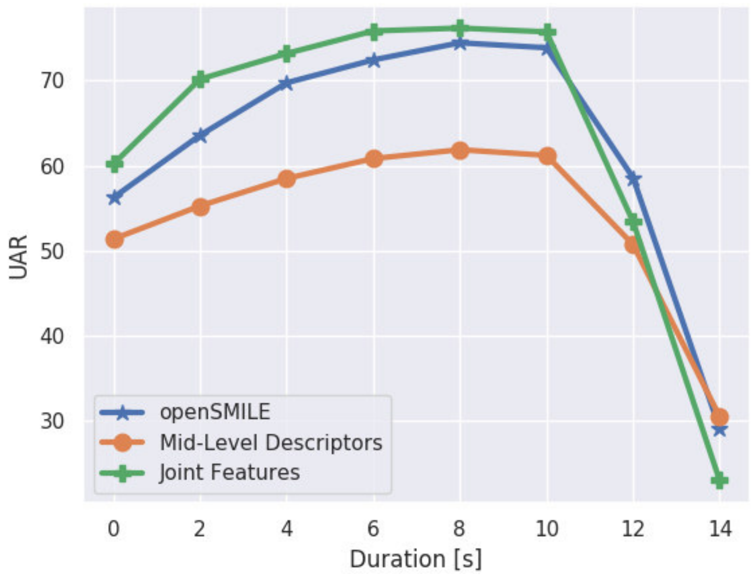

Figure 4 – UAR values achieved on the held-out test data in dependence of the feature set and the minimum duration of the analysis segments.

Table 1 – Per-language and summary evaluation metrics for all features sets and a minimum chunk duration of 8 s. kat–Georgian, kmr–Kurmanji Kurdish, pus–Pashto, tur–Turkish.

	Feature set								
	openSMILE			**MLD**			**joint**		
	precision	recall	F1	precision	recall	F1	precision	recall	F1
Language									
kat	.89	.76	.82	.81	.72	.76	.79	.89	.84
kmr	.67	.69	.68	.54	.47	.50	.62	.63	.63
pus	.78	.81	.80	.66	.66	.66	.86	.78	.82
tur	.64	.79	.71	.46	.65	.54	.77	.64	.70
Summary									
macro	.74	.76	.75	.62	.62	.62	.76	.74	.75
weighted	.77	.76	.75	.66	.64	.64	.77	.77	.76
UAR (%)	74.9			62.6			**76.3**		
accuracy (%)	76.3			64.7			76.2		

6 Discussion and conclusions

Feature set comparison Overall, higher language identification performance is achieved for openSMILE functionals than MLDs. One reason is the large difference in feature set size, 1170 functionals as opposed to 174 MLDs. Another reason is that MLDs might be more vulnerable to low audio quality. Especially MLDs derived from F0 values are affected by a less reliable F0 extraction for the given bandpass-filtered telephone speech signals. In turn, condensing the large amount of LLD functionals to a much lower-dimensional MLD vector might be beneficial for sparse data scenarios.

Minimum chunk duration For both, openSMILE functionals and MLDs, a minimum chunk duration of 8 s yields the highest performance values. This indicates that functionals and MLDs can only be reliably extracted from segments of a certain length. Especially for MLDs, this length constraint will likely differ across feature subsets. Whereas intonation-related features are best extracted from intonation phrase units, cepstral spread features require larger speech units that cover the entire vowel inventory of a language. This varying segment length dependency will be taken care of in subsequent studies.

Interpretability A solely MLD-based spoken language identification cannot compete with state of the art models achieving accuracies above 90 %. However, as illustrated in section 4, MLDs provide useful insights in language characteristics and allow for a phonetically more informed approach to language identification, potentially for a late fusion with other classifiers.

7 Acknowledgments

This work is carried out within the project *Audiobasierte Herkunftsland-Erkennung von Migranten (AUDEO)* in cooperation with the German federal police and the Hochschule für Medien, Kommunikation und Wirtschaft. This project is funded by the German Federal Ministry of Education and Research (BMBF), grant number 13N15068, in the course of the announcement "Anwender-Innovativ" in the context of the program "Forschung für die zivile Sicherheit" of the German government.

References

[1] SUGIYAMA, M.: *Automatic language recognition using acoustic features*. In *Proc. Acoustic Speech Signal Processing*, pp. 813–816. 1991.

[2] MATEJKA, P., P. SCHWARZ, J. CERNOCK, and P. CHYTIL: *Phonotactic language identification using high quality phoneme recognition*. In *Proc. Interspeech*, pp. 2237–2240. 2005.

[3] SAFITRI, N., A. ZAHRA, and M. ADRIANI: *Spoken language identification with phonotactics methods on Minangkabau, Sundanese, and Javanese language*. *Procedia Computer Science*, pp. 182–187, 2016.

[4] NG, R., C.-C. LEUNG, T. LEE, B. MA, and H. LI: *Prosodic attribute model for spoken language identification*. In *Proc. ICASSP*, pp. 5022–5025. 2010.

[5] MARTÍNEZ, D., L. BURGET, L. FERRER, and N. SCHEFFER: *Ivector-based prosodic system for language identificaton*. In *Proc. ICASSP*, pp. 4861–4864. 2012.

[6] TORRES-CARRASQUILLO, P., E. SINGER, M. KOHLER, R. GREENE, D. REYNOLDS, and J. DELLER: *Approaches to language identification using Gaussian mixture models and shifted delta cepstral features*. In *Proc. ICSLP*, pp. 89–92. 2002.

[7] CAMPBELL, W., J. CAMPBELL, D. REYNOLDS, E. SINGER, and P. TORRES-CARRASQUILLO: *Support vector machines for speaker and language recognition*. *Computer Speech and Language*, 20, pp. 210–229, 2006.

[8] LOPEZ-MORENO, I., J. GONZALEZ-DOMINGUEZ, O. PLCHOT, D. MARTINEZ, J. GONZALEZ-RODRIGUEZ, and P. MORENO: *Automatic language identification using deep neural networks*. In *Proc ICASSP*, pp. 5337–5341. 2014.

[9] JIANG, B., Y. SONG, S. WEI, J.-H. LIU, I. MCLOUGHLIN, and L.-R. DAI: *Deep bottleneck features for spoken language identification*. *PLOS ONE*, 9(7), pp. 1–11, 2014.

[10] WARD, N.: *Prosodic patterns in English conversation*. Cambridge University Press, 2019.

[11] BILLS, A.: *IARPA Babel Georgian Language Pack IARPA-babel404b-v1.0a*. Web Download. Philadelphia: Linguistic Data Consortium, 2016.

[12] BILLS, A.: *IARPA Babel Kurmanji Kurdish Language Pack IARPA-babel205b-v1.0a LDC2017S22*. Web Download. Philadelphia: Linguistic Data Consortium, 2017.

[13] ADAMS, N.: *IARPA Babel Pashto Language Pack IARPA-babel104b-v0.4bY LDC2016S09*. Web Download. Philadelphia: Linguistic Data Consortium, 2016.

[14] ANDRESEN, J.: *IARPA Babel Turkish Language Pack IARPA-babel105b-v0.5 LDC2016S10*. Web Download. Philadelphia: Linguistic Data Consortium, 2016.

[15] SCHULLER, B., S. STEIDL, A. BATLINER, J. HIRSCHBERG, J. BURGOON, A. BAIRD, A. ELKINS, Y. ZHANG, E. COUTINHO, and K. EVANINI: *The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language*. In *Proc. Interspeech*, pp. 2001–2005. San Francisco, CA, 2016.

[16] REICHEL, U.: *Unsupervised extraction of prosodic structure*. In J. TROUVAIN, I. STEINER, and B. MÖBIUS (eds.), *Elektronische Sprachverarbeitung 2017*, vol. 86 of *Studientexte zur Sprachkommunikation*, pp. 262–269. TUDpress, Dresden, Germany, 2017.

[17] LINDBLOM, B.: *Explaining phonetic variation: A sketch of the H&H theory*. In W. HARDCASTLE and A. MARCHAL (eds.), *Speech Production and Speech Modelling*, pp. 403–439. Kluwer Academic Publishers, 1990.

[18] REICHEL, U. and K. MÁDY: *Comparing parameterizations of pitch register and its discontinuities at prosodic boundaries for Hungarian*. In *Proc. Interspeech 2014*, pp. 111–115. Singapore, 2014.

[19] PIERREHUMBERT, J. and J. HIRSCHBERG: *The meaning of intonational contours in the interpretation of discourse*. In P. COEHN, J. MORGAN, and M. POLLACK (eds.), *Intentions in communication*, pp. 271–311. MIT Press, Cambridge, 1990.

[20] REICHEL, U.: *CoPaSul Manual – Contour-based parametric and superpositional intonation stylization*. RIL, MTA, Budapest, Hungary, 2016. `https://arxiv.org/abs/1612.04765`.

[21] LEHISTE, I.: *Isochrony reconsidered*. *J. Phonetics*, 5(3), pp. 253–263, 1977.

[22] SKOPETEAS, S., C. FÉRY, and R. ASATIANI: *Word order and intonation in Georgian*. *Lingua*, 119(1), pp. 102–127, 2009.

[23] VICENIK, C. and S.-A. JUN: *An autosegmental-metrical analysis of Georgian intonation*. In S.-A. JUN (ed.), *Prosodic Typology II. The Phonology of Intonation and Phrasing*, pp. 154–186. Oxford University Press, Oxford, 2014.

[24] VALLDUVÍ, E.: *The role of plasticity in the association of focus and prominence*. In *Proc. Eastern States Conference on Linguistics 7*, pp. 295–306. 1991.