Building a Persian-English OMProDat Database Read by Persian Speakers

Mortaza Taheri-Ardali, Daniel J. Hirst

▶ **To cite this version:**

Mortaza Taheri-Ardali, Daniel J. Hirst. Building a Persian-English OMProDat Database Read by Persian Speakers. Speech Prosody 2022, May 2022, Lisabon, Portugal. pp.440-444, 10.21437/SpeechProsody.2022-90 . hal-04534884

HAL Id: hal-04534884
https://hal.science/hal-04534884

Submitted on 5 Apr 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Building a Persian-English OMProDat Database Read by Persian Speakers

Mortaza Taheri-Ardali[1,2], *Daniel Hirst*[3]

[1]Department of English, Shahrekord University, Shahrekord, Iran;
[2]Bakhtiari Studies Research Center, Shahrekord University, Shahrekord, Iran
[3]Aix-Marseille University, CNRS, LPL UMR 7309, 13100, Aix-en-Provence, France
taheri@sku.ac.ir, daniel.hirst@lpl-aix.fr

Abstract

OMProDat is an open multilingual prosodic database, which aims to collect, archive and distribute recordings and annotations of directly comparable data from different languages. As part of the OMProDat project, this paper focuses on the creation of a bilingual Persian-English prosodic database read by native speakers of Persian. This collection contains 40 continuous, thematically connected paragraphs, each of five sentences, originally created during the European SAM project. Our collection was recorded by 5 male and 5 female speakers of standard Persian, all from monolingual families. The Persian texts were Romanised and transcribed phonetically using the ASCII phonetic alphabet SAMPA. The database includes TextGrid annotations, which will be obtained semi-automatically from the sound and the orthographic transcription using the SPPAS alignment software. The Momel and INSINT algorithms will be used to provide prosodic annotation of the corpus. This considerable amount of data will allow us to compare the production of Persian and English as L1 and L2, respectively. In addition, a cross-linguistic comparison with other languages in OMProDat is easily feasible.

Index Terms: Persian, speech prosody, database, OMProDat

1. Introduction

The creation of prosodic databases containing material from different languages could be particularly useful to help us to understand the complexities of human speech, especially when parallel speech data are provided. OMProDat, an open multilingual prosodic database, aims to contribute to this task [1]. The ultimate goal of the database is to collect, archive and distribute recordings and annotations of directly comparable data from an illustrative sample of different languages representing different typological prosodic characteristics. Several languages have already been incorporated into the collection so far, as described below in §2.

Modern Standard Persian was not yet included in this large parallel database. As a Southwestern Iranian language in the Indo-Iranian subdivision of the Indo-European languages, Persian has been the dominant language of Iranian lands and adjacent regions for over a millennium [2]. While Persian is the official language of Iran, it used to be the mother tongue of only about 50 percent of the population. With the spread of mass education, however, an increasing percentage of the population of Iran today speaks Persian as the first language [3].

There have been independent projects on building prosody related databases for Persian. [4], [5] and [6] are three databases made for the Persian language. [4] under the name *Persian ESD (Emotional Speech Database)* was designed for 90 sentences classified into 5 basic emotional categories. Each sentence was spoken by just one male and one female speaker. [5] is a read speech database which was designed and built specifically for Persian text-to-speech systems, taking into account Persian prosodic structure. This collection, containing 2826 phonetically and prosodically rich utterances, was recorded under studio conditions with a female voice talent speaker. The *Sharif Emotional Speech Database* or (*ShEMO*) [6] includes 3000 semi-natural utterances, extracted from online radio plays. It covers speech samples of 87 native-Persian speakers for five basic emotions plus the neutral state. Persian ESD and ShEMO are available free of charge for researchers but [5] is not freely available since it has been used in the development of the commercial Ariana Persian text-to-speech system.

The need for a systematic and cross-linguistic investigation of prosodic parameters necessitates building a new collection of natural speech data in Persian. This paper, as part of the larger project OMProDat, reports on the design and construction of a new database comprising Persian and L2 English texts read by native speakers of Persian.

2. Texts and Recordings

The text of the database is a series of 40 continuous and thematically connected five sentence passages originally created as a deliverable of the European Esprit project 2589: *SAM (Speech Assessments and Methodology)* under the name of *Eurom1* [7]. The passages were based on identical themes for the different languages, freely translated and adapted from the original English texts for the various languages. The passages were originally recorded for 11 European languages, Danish, Dutch, English, French, German, Greek, Italian, Norwegian, Portuguese, Spanish and Swedish, each speaker reading from 10 to 20 of the passages, depending on the language so that there were only two or three recordings of each passage for most languages. The original recordings were protected by copyright assigned to the different laboratories that produced the recordings.

In order to provide a more solid basis for the analysis of speech prosody, it was decided to build an open multilingual prosodic database (OMProDat), to be archived and distributed by the recently created Speech and Language Data Repository (http://sldr.org) under an open database license with new recordings for each language and with all 40 passages recorded by 10 speakers each, 5 male and 5 female [1].

The first language recorded under these conditions was Korean [8] which was recorded by 10 native speakers reading all 40 passages. This was followed by English and French,

each read by 10 native speakers as well as English read by the native speakers of French and French read by the native speakers of English [9]. This was then followed by English and Mandarin Chinese read by native speakers of Chinese [10] and Cantonese read by native speakers [11]. A number of other languages are being recorded in the same conditions.

For our new Persian recordings, all 40 passages were translated into Persian and proofread several times by the first author. We followed a free style of translation to Persianize the text as much as possible. This process helps the text to be more natural and easy to read by the speakers. As can be seen from the English gloss of the following sample text (T07), we replaced the English proper names with popular Persian equivalents.

Passage: T07

I'm trying to contact Mr. and Mrs. W. George of Swindon. They've moved from 63 Spruce Close to another part of Swindon. Can you give me their new number please? They moved approximately 3 months ago. As far as I know they're not ex-directory.

من در تلاشم تا با خانم و آقای جعفری از کاشانی تماس برقرار کنم. آنها از کوچهٔ شماره ۱۲ به کوچه ۲۴ از این خیابان نقل‌مکان کردند. می‌توانید شمارهٔ جدید ایشان را به من بدهید؟ آنها تقریباً سه ماه پیش نقل‌مکان کردند. تا آنجا که می‌دانم شمارهٔ آنها در دسترس است.

man	*dar*	*talash-am*	*ta*	*ba*	*khanom*	*va*
I	in	try-be.1sg	until	with	lady	and

aqa=ye	*jafari*	*az*	*kashani*	*tamas*	*barqarar*	*kon-am*
sir=EZ	[PN]	from	[PN]	contact	establish	do.NPST-1sg

anha	*az*	*kutshe=ye*	*shomare*	*davazdah*
they	from	alley=EZ	number	twelve

be	*kutshe=ye*	*bist=o*	*tshahar*
to	alley=EZ	twenty=and	four

az	*in*	*khiyaban*	*naqlemakan*	*kard-and*
from	this	street	movement	do.PST-3pl

mi-tavan-id		*shomare=ye*	*jadid=e*	*ishan*
IPFV-be	able.NPST-2pl	number=EZ	new=EZ	they

ra	*be*	*man*	*be-dah-id*
OBJ	to	I	IMPV-give.NPST-2pl

anha	*taqriban*	*se*	*mah=e*	*pish*
they	approximately	three	month=EZ	ago

naqlemakan	*kard-and*	*ta*	*anja*	*ke*	*mi-dan-am*
movement	do.PST-3pl	until	there	COMP	IPFV-know.NPST-1sg

shomare=ye	*anha*	*dar*	*dastres*	*ast*
number=EZ	they	in	access	be.NPST.3sg

After the preparation of the Persian translation, we provided the orthographic (Romanised) and phonemic transcriptions of the texts using the ASCII phonetic alphabet SAMPA. SAMPA (SAM Phonetic Alphabet) is a computer-readable phonetic script using 7-bit printable ASCII characters, based on the International Phonetic Alphabet (IPA). As an example, the orthographic and phonemic transcriptions of passage T07 are given below, respectively:

Passage: T07

man dar talasham ta ba khanom va aqaye jafari az kashani tamas barqarar konam anha az kutsheye shomare davazdah be kutsheye bisto tshahar az in khiyaban naqle makan kardand mitavanid shomareye jadide ishan ra be man bedahid anha taqriban se mahe pish naqle makan kardand ta anja ke midanam shomareye anha dar dastres ast

m A n$d A r$t A l a S A m$t a$b a$x a n o m$v A$? a G\ a j e$dZ A ? f A r i$? A z$k a S a n i$t A m a s$b A r G\ A r a r$k o n A m$? a n h a$? A z$k u tS e j e$S o m a r e$d A v a z d A h$b e$k u tS e j e$b i s t o$tS a h a r$? A z$? i n$x i j a b a n$n A G\ l e$m A k a n$k A r d A n d$m i t A v a n i d$S o m a r e j e$dZ A d i d e$? i S a n$r a$b e$m A n$b e d A h i d$? a n h A$t A G\ r i b A n$s e$m a h e$p i S$n A G\ l e$m A k a n$k A r d A n d$t a$? a n dZ a$k e$m i d a n A m$S o m a r e j e$? a n h a$d A r$d A s t r e s$? A s t

For the phonemes, we transcribed what the speaker actually said, not what was in the written text. For example, some sounds are not pronounced like “t” in the word “dastres” in the last line, which is just /d a s r e s/ on the Phoneme tier while the full word “dastres” is given on the Word tier.

3. Subjects

We recruited 10 native Persian speakers from Shahrekord University campus in the city of Shahrekord, Chahar Mahal va Bakhtiari Province. 5 male and 5 female speakers aged from 18-34; they all were BA English students at the time of the recordings in autumn 2018. The subjects were all born and raised in Iran. They started learning English at secondary school and enhanced their English knowledge during their BA. They were all fluent in English and were paid for their active participation in the process of sound recording. Moreover, each speaker filled in a consent form to allow us to make the data available online.

4. Recording

Recording sessions were conducted in a quiet room at Shahrekord University. They were carried out using a Shure SM58 vocal cardioids microphone (44.1 kHz, mono channel, 16-bit) connected to an Olympus LS-14 sound recorder. The microphone was set at a distance of 15 centimeters from the mouth of the speakers. It took around 2 hours for each speaker, a total of 20 hours of recording sessions. Before the recording, both English and Persian texts were given to the speakers to practice in advance. They were also asked to read the texts at a normal speaking rate and with a natural intonation. For any mistakes or long hesitations during the reading, they were asked to read the whole passage again. The total duration of the recordings for both Persian and L2 English is 5 hours and 30 minutes.

4.1. Sound files

The complete set of 40 passages for each speaker was recorded as a single sound file, including errors and repeated passages when necessary. The recordings were analysed using the Praat software [12]. A Praat TextGrid of the complete recording was made with the beginning and end of each passage labelled on an interval tier. We then wrote a Praat script to divide the original recordings into shorter files (each containing a single individual passage without errors).

The labelling convention for OMProDat is to use the 3 letter language code for the corpus defined by ISO 639-3:

[https://iso639-3.sil.org/]

Our corpus is thus named *OMProDat-pes01* since it is the first recording of speakers of Persian in the database. The L2 recordings were named *OMProDat-pes01eng*, where the first language code (*pes*) corresponds to the native language of the speakers (Persian) and the second (eng) to the language of the recordings (English), when this was different from the native language of the speaker (i.e. for the L2 recordings). The individual speakers are *pes01-f01…*, *pes01-f05*, *pes01-m01…*, *pes01-m05*. In the same vein, the recordings of each passage are *pes01-f01-t01…*, *pes01-f01-t02* etc. There is one folder for each recording which contains the individual files e.g., *pes01-f01-t01.wav*, *pes01-f01-t01.TextGrid*, etc. These folders are in the speaker folder (*pes01-f01*) which will be in the main database folder *OMProDat-pes01*.

Figure 1 shows part of the recording *pes01-f01.wav* with the accompanying hand-labelled TextGrid with a tier name Recording. Passages which were not read successfully were repeated by the speaker and only the correctly read versions were labeled in the TextGrid. The part of the recording to ignore, passages which were not read successfully, silent pauses and instructions from the recording administrator were labelled #. It's good to have a (short) silent pause at the beginning and end of each recording. In our script, the default value of this is 20 ms but this can, of course, be changed.

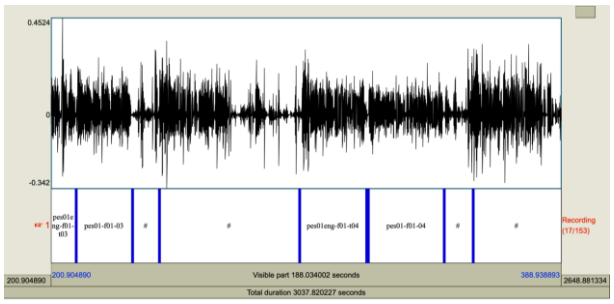

Figure 1: *Part of the TextGrid used to divide the original recording pes01-f01.wav into separate sound files, one for each passage.*

4.2. Annotations

The database contains both primary data, the recordings, and secondary data in the form of different annotation files. For the secondary data, in order to train the SPPAS aligner [13] so that it can be used for the automatic alignment of the Persian recordings, ten TextGrids were aligned by hand with the acoustic signal, with each passage being read by a different speaker. Figure 2 shows a screenshot of the manual alignment of one of the files from the Persian recordings of the database.

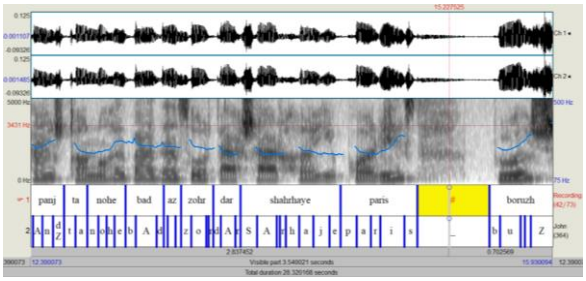

Figure 2: *The manual alignment of part of one recording i.e., T11 from the corpus* OMProDat-pes01.

5. Graphical representation

Figure 3 shows a graphical representation of the sentences "*man bayad ta saate daho si daqiqeye sobhe ruze shanbe anja basham."* (top) and "*aya mitavanid be man beguyid behtarin qatar be mashhad az masire tehran kodam ast"* (bottom) from Text 8, read by one male speaker and displayed using the OMe scale (Octave-Median) [14] which displays pitch on a logarithmic scale with the top and bottom lines corresponding to the pitch range of one octave centered on the speaker's median pitch.

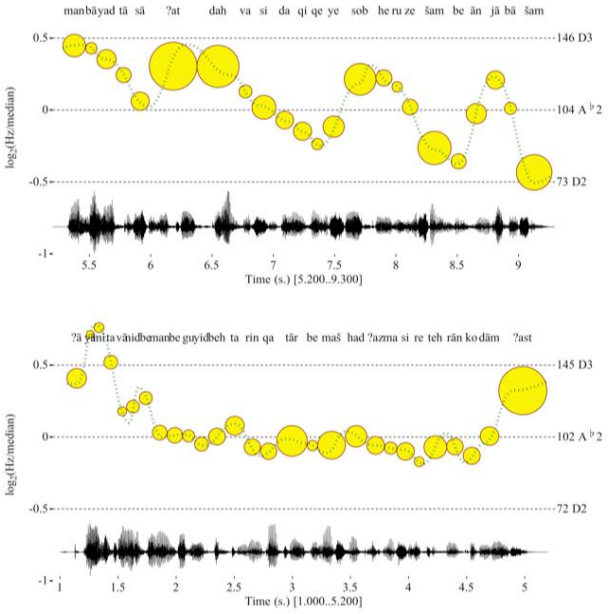

Figure 3: *Graphical representation of the sentence "*man bayad ta saate daho si daqiqeye sobhe ruze shanbe anja basham*" (top) and "*aya mitavanid be man beguyid behtarin qatar be mashhad az masire tehran kodam ast*" (bottom) by a male speaker, displayed using the OMe scale (Octave-Median).*

6. Discussion

A recent study [15] used a display like that in Figure 3, producd with the ProZed software [16] to compare the prosody of native speakers and that of Mandarin Chinese L2 speakers of English and was used to provide a visual feedback for the L2 learners to help improve their L2 prosody. The authors concluded that:

> *"audiovisual training (native sound+visual intonation contour) with a pre-class linguistic knowledge preparation and an after-class feedback is the most effective way to enhance Chinese learners' production of L2 English intonation."* [p7]

As an example of the application of this technique to our corpus, Figure 4 shows the display of the prosody of a native speaker of English (top) and that of a Persian L2 speaker of English (bottom) from our new corpus, using the OMe scale, produced with the ProZed software [16].

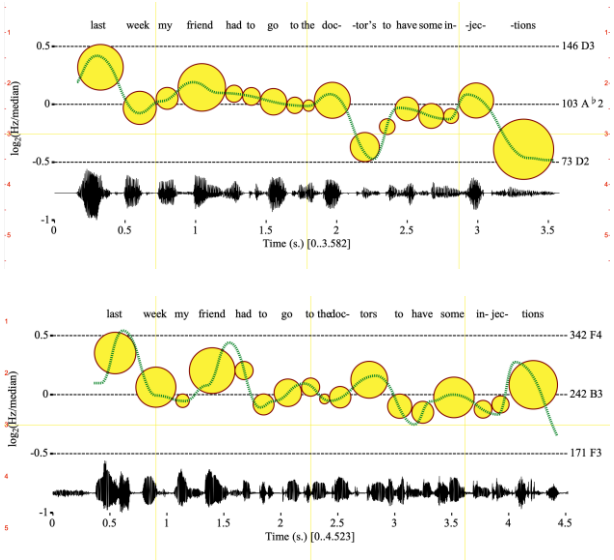

Figure 4. *Graphical display of the first sentence of the English passage T01: "*Last week my friend had to go to the doctor's to have some injections.*" read by a native speaker of English (top) and by a native speaker of Persian (bottom) using the OMe scale.*

Using the OMe scale makes it possible to compare the prosody of the native speaker (male) directly with that of the L2 speaker (female) despite the difference in pitch range between the two speakers (73 to 146 Hz and 171 to 342 Hz respectively), both of which are normalised to the pitch range of -0.5 to +0.5 octaves compared to the speaker's median pitch. The prosody of the L2 speaker in this example is actually quite good, but a comparision of the two images would make it possible for the learner to focus on the most important prosodic differences between the two readings, such as the low falling pitch on the word *doctor's* and the comparative lengthening and pitch raising of the second syllable of *injections* by the native speaker.

7. Conclusions and perspectives

This paper reported on the design and construction of a new bilingual database for Persian and L2 English, as part of the larger OMProDat Project. Our next step will be to annotate the whole database using automatic annotation tools. The automatic alignment of the acoustic signal with phonemes, syllables and words will be carried out using the SPPAS software [13], while the automatic modelling and coding of fundamental frequency will be carried out using the Momel and INTSINT algorithms [17].

Areas for future investigation are the study of melody metrics for Persian derived automatically from the acoustic signal and their comparison with those of other L1 and L2 languages in parallel to previous studies comparing English, French, Mandarin Chinese and Cantonese prosody [18, 19, 11].

We also intend to replicate the study providing visual feedback with the OMe scale [15], to compare the prosody of native and Persian L2 speakers of English in order to help native speakers of Persian to improve their L2 prosody in English.

Finally, we plan to test the usefulness of providing auditory feedback obtained by transferring ('*cloning*') the prosody of native speakers of English onto recordings by Persian L2 learners [20, 21].

All recordings and annotations will be ultimately made available online for researchers and engineers under an open-database license as part of the OMProDat database.

8. Acknowledgements

We are grateful to our Persian speakers for their contribution and their kind patience in the course of recording the database. In addition, the first author would like to thank Maryam Amani-Babadi and Leila Sadeghi-Dehcheshmeh for their assistence.

9. References

[1] D. J. Hirst, B. Bigi, H.-S. Cho, H. Ding, S. Herment, T. Wang, "Building OMProDat, an open multilingual prosodic database," in *Proceedings of TRASP, Tools and Resources for the Analysis of Speech Prosody* [satellite workshop of Interspeech], Aix-en-Provence, France, 2013, pp. 11-14.

[2] G. Windfuhr and J. R. Perry, "Persian and Tajik," in *The Iranian Languages*, G. Windfuhr, Ed. London: Routledge, 2009, pp. 416-544.

[3] G. Windfuhr and C. Jahani, "Persian," in *The World's Major Languages*, B. Comrie, Ed. London: Routledge, 2018, pp. 455-469.

[4] N. Keshtiari, M. Kuhlmann, M. Eslami, G. Klann-Delius, "Recognizing emotional speech in Persian: A validated database of Persian emotional speech (Persian ESD)," *Behav Res Methods*, vol. 47, pp. 275-294, 2015.

[5] M. Taheri-Ardali, S. Khorram, M. Assi, H. Sameti, M. Bijankhan, "Designing and recording a speech database for Persian TTS systems," vol. 6, pp. 69-84, 2016.

[6] M. O. Nezami, P. Jamshid Lou, M. Karami, "ShEMO: a large-scale validated database for Persian speech emotion detection," *Language Resources & Evaluation*, vol. 53, pp. 1-16, 2019.

[7] D. Chan, A. Fourcin, D. Gibbon, B. Granstrom, M. Huckvale, G. Kokkinakis, K. Kvale, L. Lamel, B. Lindberg, A. Moreno, J. Mouropoulos, F. Senia, I. Trancoso, C. Veld, J. Zeiliger, "Eurom - a spoken language resource for the EU," in *Proceedings of the 4th European Conference on Speech Communication and Technology*, Madrid, Spain, 1995, pp. 867-870.

[8] S. H. Kim, D. J. Hirst, H.-S. Cho, H. Y. Lee, M.-H. Chung, "Korean Multext: A Korean prosody corpus," in *Proceedings of the 4th International Conference on Speech Prosody*, Campinas, Brazil, 2008.

[9] S. Herment, A. Tortel, B. Bigi, D. J. Hirst, A. Loukina "AixOx: A multi-layered learners corpus: automatic annotation," *4th International Conference on Corpus Linguistics*, Jaèn, Spain, 2012.

[10] H. Ding and D. J. Hirst, "A preliminary investigation of third tone sandhi in Standard Chinese with a prosodic corpus," *8th International Symposium on Chinese Spoken Language Processing*, Hong Kong, 2012.

[11] D. J. Hirst, J. Wakefield, Y. H. Li, "Does lexical tone restrict the paralinguistic use of pitch? Comparing melody metrics for English, French, Mandarin and Cantonese," *International Conference on Phonetics of the Languages in China (ICPLC-2013)*, Hong Kong, 2013.

[12] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," http://www.praat.org, 1992-2021, [version 6.1.55, October 2021].

[13] B. Bigi and D. J. Hirst, "SPeech Phonetization Alignment and Syllabification (SPPAS): a tool for the automatic analysis of speech prosody," in *Proceedings of the 6th International Conference on Speech Prosody*, Shanghai, China, 2012.

[14] C. De Looze and D. J. Hirst, "The OMe (Octave-Median) scale: A natural scale for speech melody," in *Proceedings of the 7th International Conference on Speech Prosody*, Dublin, Ireland, 2014, pp. 20-23.

[15] C. Zhao, Z. Xiong, A. Li, "Using multimodal methods in L2 intonation teaching for Chinese EFL learners." in *Proceedings of the 23rd Conference of the Oriental COCOSDA International Committee for the Coordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, Yangon, Myanmar, 2020.

[16] D. J. Hirst, "ProZed: a speech prosody editor for linguists, using analysis-by-synthesis," in *Speech Prosody in Speech Synthesis. Modeling and Generation of Prosody for High Quality and Flexible Speech Synthesis*, in series *Prosody, Phonology and Phonetics*, K. Hirose and J. Tao, Eds. Berlin, Heidelberg: Springer Verlag, 2015, pp. 3-17.

[17] D. J. Hirst, "A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation," in *Proceedings of the 16th International Conference of Phonetic Sciences*, Saarbrücken, Germany, August 2007, pp. 1233-1236.

[18] H. Ding, D. J. Hirst, R. Hoffmann, "Cross-linguistic prosodic comparison with OMProDat database," in *International Conference Oriental COCOSDA & Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, 2015, pp. 212-215.

[19] D. J. Hirst, "Melody metrics for prosodic typology: comparing English, French and Chinese," in *Proceedings of Interspeech* 2013, Lyon, France, 2013, pp. 25-29.

[20] D. J. Hirst, "On the automatic comparison and cloning of native and non-native speech prosody," in *Proceedings of the 8th International Conference on Speech Prosody*, May 31-June 4, Boston, USA. 2016.

[21] D. J. Hirst, "Automatic visual and auditory feedback for second language (L2) speech prosody." [Keynote] *The 2nd International Conference on Laboratory Phonology and Phonetics (ICLPP 2)*, October 20-21, Tehran, Iran. 2021.