First Investigation of Universal Speech Attributes for Speaker Verification

Sheng Zhang[1], *Wu Guo*[1], *Guoping Hu*[2]

[1]National Engineering Laboratory of Speech and Language Information Processing,
University of Science and Technology of China, Hefei, China
[2]Key Laboratory of Intelligent Speech Technology, Ministry of Public Security, Hefei, China
zs1234@mail.ustc.edu.cn, guowu@ustc.edu.cn, gphu@iflytek.com

Abstract

The universal speech attributes to speaker verification is addressed in this paper. The manner and place of articulation form the universal attribute unit inventory, and deep neural network (DNN) is used as acoustic model. Considering the appropriate DNN output nodes, the manner and place of articulation are combined to generate more universal attribute units, which serve as DNN output nodes in acoustic modeling. The proposed attribute based DNN is used to obtain the posterior probability of the acoustic features for the total variability space training and i-vector extracting. Evaluated on the core test from the 2008 NIST speaker verification evaluation (SRE), the proposed attribute based DNN/i-vector system can achieve a comparable performance to that of the phoneme based DNN/i-vector system. Furthermore, the attribute based DNN/i-vector speaker verification system has demonstrated a good complementarity with the GMM-UBM/i-vector and phoneme based DNN/i-vector systems.

Index Terms: Speaker verification, deep neural network, universal speech attributes

1. Introduction

In recent years, i-vector [1] based speaker verification systems have become very popular for their state-of-the-art performance and ability to compensate for the channel variations. The i-vector algorithm provides a method to map a speech utterance to a low dimensional vector while retaining the speaker identity. Within this i-vector space, variability compensation methods such as linear discriminant analysis (LDA) [2] and within-class covariance normalization (WCCN) [3] are performed to reduce the channel variability. Until now, the best performance is obtained by modeling i-vector distributions through a generative model known as probabilistic linear discriminant analysis (PLDA) [4, 5, 6], which is adopted as backend classifier in this paper.

DNN has clearly shown their superiority over GMM for automatic speech recognition (ASR) [7, 8]. The methods to combine recent advances in DNN with speaker verification have attracted researchers' attention [9, 10, 11, 12, 13]. In [14], a generalized i-vector framework is proposed, where the decision tree senones (tied triphone states) of a DNN model in the ASR system are employed to generate posterior probabilities, rather than the conventional GMM-UBM. In combination with a PLDA backend, the DNN/i-vector framework can significantly improve the speaker verification performance.

The DNN/i-vector framework adopts the classical ASR acoustic model (AM) to produce frame alignments, and the AM is a language dependent model and must be trained using language specific data. As we know, there is no direct relationship between the phonetic information of a speaker's speech and characteristic of his vocal track. It is more reasonable to find fundamental units which can be defined universally across all languages. In this paper, we propose to replace phonetic information with universal speech attributes for speaker verification. There is a growing interest in exploiting the discriminative properties of universal speech attributes in speech processing [15, 16, 17]. S. M. Siniscalchi et al. adopted universal speech attributes on the token based spoken language identification (LID); promising results comparable to acoustically rich phone based LID systems have already been obtained [18]. A fusion approach is proposed to LID by combining multiple tokenizers with phone and speech attributes models to achieve an additional average relative equal error rate (EER) reduction in [19], which demonstrates that speech attribute units are complementary to phone units.

In this work, we aim to use a universal speech attribute based DNN to guide speaker modeling. The output nodes of the DNN are tied triple-attributes states. Considering the number of tied triple-attribute states, a method to combine the manner and place of articulation has been proposed to generate attributes units. The attribute based DNN is used to generate the posterior probability of acoustic features in speaker verification. After the posterior probability, namely the zeroth-order statistics, is obtained, the firstorder statistics is computed in the standard manner. The advantage of using attributes units is that they are more fundamental than phonemes, and the acoustic model can be trained using different language corpus. In other words, universal speech attributes are more related to the pronunciation habits of a person than the speech content.

The remainder of this paper is organized as follows. In section 2, we describe how to obtain universal speech attribute units in AM modeling. Then, we briefly review the DNN/i-vector system in section 3. In section 4, results using the attribute based DNN on the NIST SRE 2008 corpus are presented. Finally, we conclude our paper in section 5.

2. Universal Speech Attributes

We build a large vocabulary continuous speech recognition (LVCSR) system using the universal speech attributes, and this acoustic model is used to obtain frame alignments for speaker verification. The difference between the proposed and the classical LVCSR systems is the replacement of phonemes with attribute units in the acoustic model. The input acoustic features and training procedure of the proposed system are identical to those of the classical phoneme based system.

The set of universal speech attributes is listed in the first and second rows of Table 1, which consists of the place and manner of articulation [16]. The numbers of manner and the place of articulation are 11 and 10, respectively, which are much fewer than the phoneme set (approximately 40 in English ASR) in conventional LVCSR system. In LVCSR acoustic model training, context-dependent (CD) models are always adopted to improve the recognition accuracy. Even when the context-dependent models are used, the number of attribute units is not sufficient for good recognition performance. It is unwise to separately use the place and manner of articulation in acoustic model.

We combine the place and manner of articulation to increase the number of attribute units. Because there is a direct mapping between the phonemes and attribute units, we can use phonemes to generate more attribute units. We look up the place and manner of a phoneme. If they are different from those of other phonemes, we define a new attribute unit. For example, the manner and place of phoneme /ah/ are /vowel/ and /mid/, respectively, so we define a new attributes unit /mid_vowel/. The English phoneme set is used in our experiments, and 23 universal speech attribute units are obtained by combining the place and manner of articulation. The set of combination attributes is listed in the third row of Table 1. In addition to the listed attributes set, the /silence/ token is used to represent the soundless segments, the /stop/ token denotes pauses between speech, and the /garbage/ token represents /garbage/, /noise/, /breath/, /cough/, /laugh/, /lip smack/, /sigh/ and /sneeze/.

Table 1: *Universal Speech Attributes list in terms of the manner and place of the articulation*

Manner	affricate, fricative, nasal, vowel, voice-stop, unvoiced-stop, glide, liquid, diphthong, sibilant
place	alveolar, alveo-palatal, dental, glottal, high, bilabial, labio-dental, low, mid, palatal, velar
Place_manner	mid_vowel, alveo-palatal_affricate, alveolar_voice-stop, low_diphthong, palatal_glide, mid_diphthong, velar_unvoiced-stop, high_vowel, velar_voice-stop, alveo-palatal_sibilant, low_vowel, alveolar_unvoiced-stop, dental_fricative, labio-dental_fricative, alveolar_sibilant, high_diphthong, bilabial_voice-stop, bilabial_glide, alveolar_liquid, alveolar_nasal, bilabial_nasal, bilabial_nasal, velar_nasal, bilabial_unvoiced-stop, glottal_fricative

All of aforementioned attribute units are used in the following acoustic modeling exactly as phonemes are used in the state-of-the-art ASR systems. Furthermore, content dependent modeling is adopted to improve performance. As we lack linguistic knowledge for attribute units, we can't design a suitable question set for state tying. In this work, we generate question set using the approach described in [20]. It is a clustering technique based on likelihood maximization criteria, where the first and second half states of context independent models are used to generate right-context and left-context questions. In the clustering procedure, groups of attribute units are recursively clustered until only two maximally separated clusters, and the procedure is repeatedly performed on each of these clusters with subsequent exhaustive partitioning. The procedure stops when the number of attribute units in both maximally separated clusters is less than or equal to 2. All maximally separated clusters, which are pairwise generated in the clustering procedure, are parts of final question set. The detailed description can be referred in [20].

The training procedure of attribute units based acoustic model is identical to that of the conventional phoneme based systems, and we use DNN model to generate the posterior probability of feature vectors, which will be used in the following total variability modeling.

3. The DNN/i-vector Framework

In the i-vector model [1], we assume that the following distribution generates the t-th speech frame $\boldsymbol{x}_t^{(i)}$ from the i-th speech sample:

$$\boldsymbol{x}_t^{(i)} \sim \sum_k \gamma_{kt}^{(i)} N(\boldsymbol{\mu}_k + \boldsymbol{T}_k \boldsymbol{\omega}^{(i)}, \boldsymbol{\Sigma}_k) \tag{1}$$

where the $\boldsymbol{T}_k$ matrices describe a low-rank subspace (called total variability subspace), by which Gaussian means are adapted to a particular speech segment; $\boldsymbol{\omega}^{(i)}$ is a segment-specific standard normal-distributed latent vector; $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean and covariance of the k-th component of the UBM, respectively. In the training procedure, $\boldsymbol{T}_k$ matrices can be estimated using zeroth- and first-order Baum-Welch statistics of the training corpus.

In the GMM-UBM/i-vector framework, each utterance is represented by its zeroth- and first-order Baum-Welch statistics extracted with the UBM. In paper [14], Y. Lei et al. made an important modification to estimate the statistics. They adopted an ASR DNN model to generate the zeroth-order statistics of feature vector $\boldsymbol{o}_t$. In this paper, we use a similar DNN/i-vector framework with [14]. The only difference is our replacement of the phoneme based DNN with the proposed attribute based DNN. The flowchart of attribute based DNN/i-vector system is shown in Fig. 1. In the DNN/ i-vector framework, the UBM can be trained in a supervised fashion. The means and covariance of this UBM are:

$$\begin{aligned} \gamma_{kt}^{(i)} &\approx p(k|\boldsymbol{x}_t^{(i)}) \\ \boldsymbol{\mu}_k &= \frac{\sum_{i,t} \gamma_{kt}^{(i)} \boldsymbol{x}_t^{(i)}}{\sum_{i,t} \gamma_{kt}^{(i)}} \\ \boldsymbol{\Sigma}_k &= \frac{\sum_{i,t} \gamma_{kt}^{(i)} \boldsymbol{x}_t^{(i)} \boldsymbol{x}_t^{(i)T}}{\sum_{i,t} \gamma_{k,t}^{(i)}} - \boldsymbol{\mu}_k {\boldsymbol{\mu}_k}^T. \end{aligned} \tag{2}$$

The attribute based DNN is used to compute the posteriors $p(k|\boldsymbol{x}_t^{(i)})$ for each frame. This new supervised UBM can replace the traditional unsupervised UBM to obtain the required statistics for the i-vector computation. After the new supervised UBM is obtained, we can obtain the required posteriors for the i-vector computation. Given a speech segment i, the following sufficient statistics can be computed using the DNN posterior probabilities,

$$\begin{aligned} N_k^{(i)} &= \sum_t \gamma_{kt}^{(i)} \\ \boldsymbol{F}_k^{(i)} &= \sum_t \gamma_{kt}^{(i)} \boldsymbol{x}_t^{(i)} \end{aligned} \tag{3}$$

where N_k and $\boldsymbol{F}_k$ represent the zeroth- and first-order statistics, respectively. These sufficient statistics are used to train the subspace $\boldsymbol{T}_k$ and extract the i-vector $\boldsymbol{\omega}^{(i)}$.

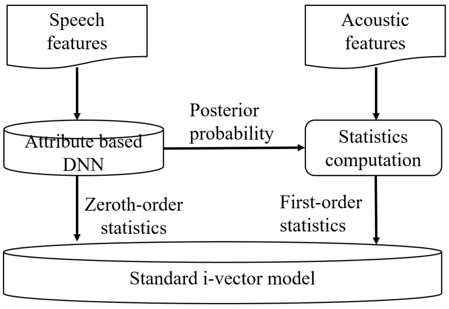

Figure 1: *The flow diagram of attribute based DNN/i-vector framework*

4. Experiments

4.1. ASR results

To obtain a fair comparison between the attribute based DNN and the conventional phoneme based DNN systems, both HMM-GMM and HMM-DNN models are trained using approximately 300 hours of clean English telephone speech from Switchboard data sets. Almost 4000 tied tri-attribute states are obtained using decision trees as described in [20]. Standard HMM-GMM systems are used to generate the initial states alignments to train these two DNNs. Except for the output layer, these two DNNs have identical architectures. The inputs of DNNs are 429-dimensional features, corresponding to 39-dimensional perceptual linear predictive (PLP) features within a context window of 11 (5+1+5) frames. There are 6 hidden layers with 2048 hidden units in each layer. The output layer of phoneme based DNN has 3990 units, whereas that of the attribute based DNN has 3988 units. To make the characteristics conform to a Gaussian distribution, the features are preprocessed with mean and variance normalization (MVN). The cross entropy criterion is used to train the DNN model.

Because the DNNs are only used in posteriors extracting, we do not need to compare the ASR word accuracy. A frame classification experiment is conducted on the Hub5e00 corpus. The frame accuracy of these two DNNs is listed in Table 2.

Table 2: *Frame accuracy on Hub5e00*

Acoustic model	Accuracy (%)
Phoneme based DNN	45.17
Attribute based DNN	43.80

Table 2 shows that the frame accuracy is not very high for both the phoneme and attribute based DNNs. The phoneme based DNN outperforms the attribute based DNN by 1.37%, and this performance gap is not notably obvious for the speaker verification.

4.2. Speaker verification results on NIST 2008

The experiments are carried out under the common conditions 6, 7 and 8 of the NIST 2008 SRE database. The training and test conditions of these three common conditions are as followed:

- C6: All trials involving only telephone speech in training and test.
- C7: All trials involving only English language telephone speech in training and test.
- C8: All trials involving only English language telephone speech spoken by a native U.S. English speaker in training and test.

The 39-dimensional PLP features are used in the experiments. Each speech signal is parameterized by the 13th order PLPs and their first and second derivatives. Further processing including relative spectral (RASTA) filtering, voice activity detector (VAD), cepstral mean subtraction (CMS) and gaussianization are applied to all PLPs.

Three i-vector systems, including the GMM-UBM/i-vector, the phoneme based DNN/i-vector and the proposed attribute based DNN/i-vector systems, are compared in this paper. Because these three systems include similar procedure in training procedure, the same file lists from previous SRE databases are selected as training set. NIST SRE 2004, 2005, 2006 and switchboard corpora are used to train the UBMs. For the baseline GMM-UBM/i-vector system, Gender-dependent UBMs with 1024 components are trained using the expectation maximization (EM) algorithm. For the DNN based systems, supervised UBMs are trained through the aforementioned DNN posteriors. Each DNN output node is modeled by a single Gaussian. All Gaussian components are merged into a UBM model for the following i-vector model training.

After the UBM model is obtained, the conventional total variability matrix training and i-vector extraction procedures are performed. The total variability matrix with rank 400 is trained using the NIST SRE databases before 2008. After extracting the i-vector, further processing including LDA, WCCN, whitening and length normalization algorithms are applied to improve the performance. PLDA algorithm is used as backend classifier, where the sizes of speaker and channel matrices are 150 and 10, respectively.

The EER and minimal detection cost function (DCF) are used to evaluate the performance of the systems. The performances of different systems (Sys1 to Sys3) are listed in Table 3. For comparison, the phoneme and attribute based DNN models are both trained on English telephone speech from Switchboard data sets. From Table 3, the traditional GMM-UBM/i-vector system (Sys1) achieves the best performance in multilingual condition (i.e., C6), whereas the supervised methods (sys2 and sys3) outperform Sys1 in language matched conditions (i.e., C7 and C8). A reasonable explanation is that the DNNs can provide more accurate posteriors than the unsupervised GMM-UBM in language matched conditions. Furthermore, the attribute based DNN/i-vector achieves comparable performance to that of the phoneme based DNN/i-vector. There is a slight performance gap because of the relatively rough modeling unit of the attribute based DNN.

As an important merit, the universal speech attributes can get rid of the restrictions on the language, and the attribute based DNN can be trained on multilingual corpus to improve the performance. In this section, 140-hour Mandarin telephone speech files are added to the original 300-hour Switchboard training corpus to train the universal speech attributes recognizer. We map Mandarin phoneme to universal speech attributes as we has done for English. After this procedure, we use aforementioned data-driven method to generate decision trees, and 3993 tied tri-attribute states are obtained. Except for the output layers, the identical architectures as Sys3 are adopted to train the

Table 3: *Experimental results in NIST SRE (EER% / minDCF08*1000)*

System description	Training speech	female			male		
		C6	C7	C8	C6	C7	C8
Sys1: GMM-UBM/i-vector	–	**5.68/28.9**	2.54/12.8	2.91/13.3	**3.73/20.5**	1.65/9.04	1.09/5.47
Sys2: Phoneme based DNN	English	6.13/31.6	**1.82/9.56**	**1.93/9.51**	4.79/21.1	**1.44**/8.22	**0.64**/3.82
Sys3: Attribute based DNN	English	6.79/34.2	1.93/11.1	2.16/10.9	5.12/22.6	2.03/**7.47**	**0.64**/4.26
Sys4: Attribute based DNN	English + Mandarin	6.38/33.7	1.91/10.6	**1.93**/11.0	5.11/21.2	1.55/8.10	0.87/**3.61**
Fusion: sys1+sys2	–	5.28/27.7	1.78/10.0	2.12/10.2	3.84/**18.4**	1.55/7.63	0.83/3.72
Fusion: sys2+sys3	–	6.10/31.5	1.72/9.69	1.81/9.87	4.76/21.1	1.82/7.67	**0.52**/3.82
Fusion: sys2+sys4	–	6.05/32.1	1.78/9.49	1.88/9.23	4.59/20.7	1.50/**7.12**	0.79/3.60
Fusion: sys1+sys3	–	5.51/29.0	1.85/10.8	1.89/11.0	3.86/19.0	1.78/7.86	0.85/4.48
Fusion: sys1+sys4	–	5.34/28.7	1.81/10.5	1.90/10.7	3.76/18.8	1.55/7.42	0.88/3.94
Fusion:sys1+sys2+sys3	–	**5.16**/27.9	**1.71/9.02**	1.82/**9.11**	**3.50**/19.1	1.35/7.64	0.57/3.49
Fusion:sys1+sys2+sys4	–	5.21/**27.5**	1.73/9.21	**1.79**/9.17	**3.50**/18.5	**1.32**/7.26	0.57/**2.85**

attribute based DNN, and this system is denoted as Sys4. Compared with the second row of Table 2, the frame accuracy of this DNN is reduced to 42.97%. A reasonable explanation is that the characteristic of the additional Mandarin corpus does not match that of the Switchboard corpus.

After training the attribute based DNN using English and Mandarin corpus, the posteriors are generated, and speaker verification experiments are performed. The performance of Sys4 is listed in Table 3. Compared with Sys3, Sys4 can achieve recognition improvement in most conditions. Thus, the addition of Mandarin training data brings a more robust DNN model, which generates more accurate posteriors than Sys3. In particular, for C6 condition, Sys4 is consistently improved compared to Sys3 on both male and female parts. Because C6 condition is a multilingual speaker verification task, the DNN that is trained with multilingual corpus has stronger classification ability than the DNN trained with a single language corpus.

4.3. Score fusion

The score fusion of different systems is a challenging issue in speaker verification field. Because all these four systems can achieve similar recognition performance, a notably simple score fusion method is adopted in this paper. The scores of different systems are fused with equal weights. The fusion results are shown in the last seven rows of Table 3. The fusion of different systems do not always provide improved performance over the single best system, but it can perform better than the single system in most common conditions. Specifically, the attribute based DNN/i-vector speaker verification system have demonstrated a good complementarity with the GMM-UBM/i-vector and phoneme based DNN/i-vector systems on both male and female parts.

5. Conclusion

One limitation of the phoneme based DNN/i-vector framework is the lack of universal acoustic characterization. This paper presents a novel speaker verification system, where universal speech attributes are used to define a new attribute based DNN that generates frame alignments. This system opens new potentials to use universal acoustic characterization for speaker verification and train the DNN model with multilingual corpus. The attribute based DNN/i-vector system achieves comparable performance with the phoneme based DNN/i-vector system. Furthermore, the performance of the attribute based DNN/i-vector system can be considerably improved when it is trained with English and Mandarin corpora. Finally, the fusion of the attribute based DNN/i-vector system and other conventional systems can obtain an obvious improvement. The results indicate that the attribute based DNN/i-vector system and phoneme based DNN/i-vector system are essentially complementary to each other.

We design the attribute units using the phoneme mapping in section 2, and this method has direct connection with the English phoneme set. In our following study, we will continue to design a more robust set of attribute units using data-driven methods.

6. Acknowledgement

This work was partially funded by the Natural Science Foundation of Anhui Province (Grant No. 1408085MKL78).

7. References

[1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio Speech & Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[2] C. Bishop, "Pattern recognition and machine," New York:Springer, 2006.

[3] A. O. Hatch, S. S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for svm-based speaker recognition." in *Interspeech*, 2006.

[4] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "Plda for speaker verification with utterances of arbitrary duration," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7649–7653.

[5] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[6] P. Kenny, "Bayesian speaker verification with heavy-tailed priors." in *Odyssey*, 2010, p. 14.

[7] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[8] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 30–42, 2012.

[9] M. Senoussaoui, N. Dehak, P. Kenny, R. Dehak, and P. Dumouchel, “First attempt of boltzmann machines for speaker verification.” in *Odyssey*, 2012, pp. 117–121.

[10] T. Stafylakis, P. Kenny, M. Senoussaoui, and P. Dumouchel, “Preliminary investigation of boltzmann machine classifiers for speaker recognition.” in *Odyssey*, 2012, pp. 109–116.

[11] S. Yaman, J. W. Pelecanos, and R. Sarikaya, “Bottleneck features for speaker recognition.” in *Odyssey*, vol. 12, 2012, pp. 105–108.

[12] V. Vasilakakis, S. Cumani, and P. Laface, “Speaker recognition by means of deep belief networks,” 2013.

[13] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, “Deep neural networks for extracting baum-welch statistics for speaker recognition,” in *Odyssey*, 2014.

[14] Y. Lei, L. Ferrer, M. McLaren *et al.*, “A novel scheme for speaker recognition using a phonetically-aware deep neural network,” in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1695–1699.

[15] S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, “Exploiting context-dependency and acoustic resolution of universal speech attribute models in spoken language recognition,” in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[16] ——, “Universal attribute characterization of spoken languages for automatic spoken language recognition,” *Computer Speech & Language*, vol. 27, no. 1, pp. 209–227, 2013.

[17] V. Hautamäki, S. M. Siniscalchi, H. Behravan, V. M. Salerno, and I. Kukanov, “Boosting universal speech attributes classification with deep neural network for foreign accent characterization,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[18] S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, “Exploring universal attribute characterization of spoken languages for spoken language recognition.” in *INTERSPEECH*, 2009, pp. 168–171.

[19] Y. Wang, J. Du, L. Dai, and C.-H. Lee, “A fusion approach to spoken language identification based on combining multiple phone recognizers and speech attribute detectors,” in *Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on*. IEEE, 2014, pp. 158–162.

[20] R. Singh, B. Raj, and R. M. Stern, “Automatic clustering and generation of contextual questions for tied states in hidden markov models,” *Proceedings of the Icassp Phonexi Az*, vol. 1, pp. 117–120, 1999.