一般社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

発音構造分析に基づく話者を単位とした世界英語の発音クラスタリング

沈涵平[†‡] 峯松信明[‡] スティーブン・ワインバーガー[††] 牧野武彦[‡‡] ノバック・ジョセフ[‡]
ポンキッティパン・ティーラポン[‡] 呉宗憲[†]

†国立成功大学, 〒701 台湾台南市大学路 1 号
‡東京大学, 〒113-8656 東京都文京区本郷 7-3-1
††ジョージ・メイソン大学, 4400 University Dr, Fairfax, VA 22030
‡‡中央大学〒192-0393 東京都八王子市東中野 742-1

E-mail: [‡]{happy,mine,novakj,bank}@gavo.t.u-tokyo.ac.jp, [††]accent@gmu.edu, [‡‡]mackinaw@tamacc.chuo-u.ac.jp, [†]chwu@csie.ncku.edu.tw

あらまし 英語は，国際的な言語コミュニケーションを可能にする唯一の言語である。しかし，話者の出身地・成育環境に依存して，英語の発音には地方訛り・外国語訛りが不可避的に混入する。本研究の究極の目的は，世界英語を対象とした個人を単位とする発音の世界地図を作成することにある。この地図を使うことで，近しい発音をする話者を見つけることが出来，英会話相手として最適な話者も見つけることができるようになる。と同時に，自身の発音が他者とどれくらい違うのかも知る事ができる。このような地図の作成は，数学的には，対象とする全話者に対する発音距離行列を求めることが必要である。本研究では，話者不変性を持つ発音構造分析とサポートベクタ回帰を話者間の発音距離推定に応用する。世界英語データとしては，Speech Accent Archive を利用し，学習・評価データとして使用した。実験の結果，非常に精度の高い話者間の発音距離予測ができることが示された。

キーワード 世界英語，話者を単位とした発音クラスタリング，発音構造，サポートベクター回帰

Speaker-based pronunciation clustering of World Englishes based on pronunciation structure analysis

H.-P. SHEN[†‡], N. MINEMATSU[‡], S. H. WEINBERGER[††], T. MAKINO[‡‡], J. R. NOVAK,
T. PONGKITTIPHAN[‡] and[‡] C.-H. WU[†]

† National Cheng Kung University, No.1, University Road, Tainan City 701, Taiwan
‡ The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656 Japan
† † George Mason University, 4400 University Dr, Fairfax, VA 22030
‡ ‡ Chuo University, 742-1 Higashinakano Hachioji-shi, Tokyo 192-0393 Japan

E-mail: [‡]{happy,mine,novakj,bank}@gavo.t.u-tokyo.ac.jp, [††]accent@gmu.edu, [‡‡]mackinaw@tamacc.chuo-u.ac.jp, [†]chwu@csie.ncku.edu.tw

Abstract English is the only language available for global communication. Due to the influence of the students' mother tongue, however, speakers from different regions inevitably have different accents in their pronunciation of English. The ultimate goal of our project is creating a global pronunciation map of World and individual Englishes, for speakers to use to locate similar English pronunciations. The speaker can then find the best English conversation partner. A learner can also know how his pronunciation geographically compares to other varieties. Creating a map mathematically requires a matrix of pronunciation distances among all the speakers considered. This paper investigates invariant pronunciation structure analysis and SVR to predict inter-speaker pronunciation distances for new speaker pairs. The speech accent archive, containing data from worldwide accented English speech, is used as training and testing samples. Experiments show very promising results.

Keyword World Englishes, Speaker-based Pronunciation Clustering, Pronunciation Structure, Support Vector Regression

1. Introduction

English is the only language available for global communication. In many schools, native pronunciation of English is presented as a reference, which students try to imitate. It is widely accepted, however, that native-like pronunciation is not always needed for smooth communication. Due to the influence of the students' mother tongue, those from different regions inevitably have different accents in their pronunciation of English. Recently, more and more teachers accept the concept of World Englishes [1,2] and they regard US and UK pronunciations as just two major examples of accented English. If one takes the philosophy of World Englishes as it is, we can claim that every kind of accented English is equally correct and equally incorrect. In this situation, there is a great interest in how one type of pronunciation is different from another. As shown in [3], the intelligibility of spoken English heavily depends on the nature of the listeners, and foreign accented English can indeed be more intelligible than native English. Generally speaking, intelligibility tends to be enhanced among speakers of similarly accented pronunciation.

The ultimate goal of our project is creating a global map of World and individual Englishes, for speakers to use to locate similar Englishes. The speaker can then find the best English conversation partner. A learner can also know how his pronunciation geographically compares to other varieties. If he is too distant from these other varieties, he may have to correct his pronunciation for the first time to achieve smoother communication with these others. For this project, we have two major problems. One is collecting data and labeling them, and the other is creating a good algorithm of drawing the global map. Luckily enough, for the first problem, the third author has made a good effort in systematically collecting World Englishes from more than a thousand speakers from all over the world. This corpus is called Speech Accent Archive [4]. To solve the second problem in this paper, we propose a method of clustering speakers only in terms of their pronunciation. Clustering of items can be done by calculating a distance matrix among them. The technical challenge here is how to calculate the pronunciation distance between any pair of the speakers in the archive, where irrelevant factors involved in the data, such as differences in age, gender, microphone, channel, background noise, etc have to be ignored adequately. For that, we use a pronunciation structure paradigm [5,6] with support vector regression (SVR). Our experiments demonstrate very promising results.

2. Speech Accent Archive

The corpus is composed of read speech samples of more than 1,700 speakers and their corresponding IPA transcriptions. The speakers are from different countries around the world and they read a common elicitation paragraph, shown in Fig. 1. It contains 69 words and can be divided into 221 phonemes using the CMU dictionary as reference [7]. Each sample has its detailed IPA transcription, which is provided by trained phoneticians, and an example is also shown in Fig. 1. The transcriptions can be used to prepare a reference for inter-speaker distances, which will be adopted as a target of prediction using support vector regression in our study.

The recording condition in the corpus varies among samples because the audio data were collected under many different situations. To create a suitable map, these acoustic variations including age- and gender-variation have to be cancelled well because these are totally irrelevant to clustering the speakers in terms of pronunciation.

In this study, only the data with no word-level insertion or deletion were used. The audio files with exactly 69 words were selected as candidate files and 515 speakers' files were obtained. Some of these files were found to include a very high level of background noise, and we manually removed them. At the end of the day, 381 speakers' data were obtained and used in our study.

"Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station."

[pʰl̥iːz kʰɑːl stɛlʌ æsk hɚ ɾə bɹĩŋ ðiːz θĩŋz wɪθ hɛɪ fɪʌ̃m ðə stoːɹ sɪks spʊ̃ːnz əɣ fɪɛʃ snoʊ pʰiːʐ faːɣ θɪkˀ slæːbz ə blụː ʧĩːʐ ɛ̃n meɪbi ə snækˀ foɪ hɚ bɹʌðə bɑb wi ɑlso n̪iːd ə smɑːl pʰl̥æstɪk sneɪkˀ æ̃n ə bɪgˀ tʰɔɪ fɪog fõɪ ðə kʰɪdʐ ʃi kʰĩn skụːp ðiːʐ θĩŋz̃ ĩntʰụ θɪi ɪɛ̨d bæːgz æ̃n wi wɪl miːt hɛɪ wɛ̃ntsdi æt ðə tɹeĩn steɪʃĩn]

Fig. 1 The elicitation paragraph and an example of IPA transcription

3. Inter-speaker Pronunciation Distance

A pronunciation distance predictor based on pronuncia-

tion structure was constructed. To this end, we prepared reference inter-speaker distances in the speech data, which can be used to train the distance predictor and verify the predicted distances. In this paper, the reference pronunciation distance between two speakers is calculated through comparing their individual IPA transcriptions. Since all the transcriptions contain exactly the same number of words, word-level alignment between transcriptions is easy and we only have to deal with phone-level insertions, deletions, and substitutions between a word and its counterpart. It should be noted that diacritical marks in our IPA transcriptions were ignored because we wanted to focus mainly on phone-level differences between the two transcriptions. DTW-like comparison between a word and its counterpart gives us a penalty score depending on what kind of phone-level changes are found between the two. For insertion and deletion, a high penalty is given because they change syllable structure**.** For substitution, a lower penalty is assigned. Furthermore, in [4], some phonological generalization rules, which were defined by phoneticians, are used to describe each speaker's pronunciation. These rules represent commonly observed substitution patterns. As they are very common, the lowest penalty was assigned to them. The phonological generalization rules, penalties and their corresponding examples are shown in Table 1. By accumulating these phone-level penalties, a word-based penalty score was obtained. By accumulating these word-level penalties, we get the paragraph-based penalty between two speakers. Moreover, by normalizing the penalty using the averaged number of phones over the two transcriptions, the final (and normalized) penalty score was obtained. This was defined as the inter-speaker pronunciation distance in this study.

Although the final and normalized scores are used as reference for inter-speaker distances in the following sections, we do not claim at all that the above procedure of calculating scores is the best and only procedure for our purpose. Our definition of penalty scores is very heuristic and what we want to claim here is that our proposed "prediction" algorithm is expected to work independently of the definition of penalties.

Table 1 The used phonological generalization rules and penalties in calculating reference inter-speaker distances

Phonological generalization rules	Examples	Penalty (Distance increasing)
Final obstruent devoicing & Consonant voicing	b ⇔ p t ⇔ d	+1
Stop (plosive) =>Fricative	p => ɸ b => β	+1
Interdental fricative change	θ => t θ => d	+1
Alveolar approximant change	ɹ => r ɹ => X	+1
w => fricative	w => v	+1
h => velar fricative	h => x h => ɣ	+1
S -> s	ʃ => s	+1
Syllable structure change	Vowel insertion, Consonant deletion, Consonant insertion	+5
Phone-level substitution	Other substitutions	+3

4. Invariant Pronunciation Structure

Minematsu et al. proposed a new method of representing speech, called speech structure, and proved that the acoustic variations, corresponding to any linear transformation in the cepstrum domain, can be completely unseen in the representation [5]. This invariance is attributed to the invariance of Bhattacharyya distance (BD), which is calculated using equation 1 and is proved to be invariant with any linear transform.

$$D_B = \frac{1}{8}(\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2) + \frac{1}{2}\ln\left(\frac{\det\Sigma}{\sqrt{\det\Sigma_1 \det\Sigma_2}}\right) \qquad (1)$$

where μ_1, μ_2 are mean vectors and Σ_1, Σ_2 are covariance matrices of two Gaussian distributions. $\Sigma = (\Sigma_1+\Sigma_2)/2$.

By calculating the BD of every pair of sound units in the elicitation paragraph read by a specific speaker, the unique distance matrix with respect to that speaker can be obtained. This sound structure is called the pronunciation structure in this paper. The structure only represents the local and global contrastive aspects of a given utterance, which is theoretically similar to Jakobson's structural phonology [8]. [7] showed experimentally that the invariant pronunciation structure is useful to group dialects into clusters. Thus, the differences of the structures between two speakers can be used as features to estimate inter-speaker pronunciation distances.

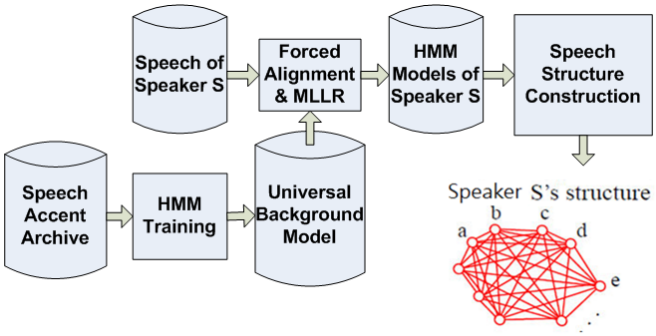

Fig.2 Speaker-dependent Pronunciation Structure

Fig. 2 shows the process to construct pronunciation structure. To construct a specific speaker's pronunciation structure, we first trained a paragraph-based universal background hidden Makov model (HMM) using all the data available. Conventional 24-dimensional MFCCs (MFCC +Δ MFCC) were used to train the HMM. Here, the paragraph was converted into a phoneme sequence using the CMU dictionary and, by using this as reference phoneme string, a paragraph-based HMM was built using all the data. Then, for each speaker, forced alignment of that speaker's utterance was done to obtain phoneme boundaries and MLLR adaptation was done to adapt the universal model to that speaker. In MLLR adaptation, the number of regression classes used was 32. In this adapted model, a phonemic segment was characterized as three states. Each state contains one Gaussian. Finally, three BDs are calculated between a phonemic segment and another in an input utterance. They are averaged and rooted to give us the final score ($d_{p_i p_j}$) between the two phonemic segments in equation 2.

$$d_{p_i p_j} = \sqrt{\frac{BD(p_i^1, p_j^1) + BD(p_i^2, p_j^2) + BD(p_i^3, p_j^3)}{3}} \qquad (2)$$

where p_i and p_j denote the i-th and j-th phone, respectively. p^1, p^2 and p^3 are the first, second and third states of the phonemic segment p. All the distances $d_{p_i p_j}$ are used together to derive the pronunciation structure. The distance matrix S_{matrix} of specific speaker S can be represented as equation 3.

$$S_{matrix} = \begin{bmatrix} 0 & d_{p_1p_2} & . & . & . & d_{p_1p_N} \\ d_{p_2p_1} & 0 & . & . & . & . \\ . & . & 0 & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & 0 & d_{p_{N-1}p_N} \\ d_{p_Np_1} & . & . & . & d_{p_Np_{N-1}} & 0 \end{bmatrix} \qquad (3)$$

In the distance matrix, the diagonal elements are all zero because the distance from a phonemic segment to itself is zero. $d_{p_i p_j}$ and $d_{p_j p_i}$ are the same because the distance is estimated from the same phone pair. Only the elements found in the upper triangle are used to form the pronunciation structure of a specific speaker. The construction process of sentence-based pronunciation structure is shown in Fig. 3.

For two given pronunciation structures (two distance matrices) from speakers S and T, a difference matrix between the two is calculated by equation 4, which is shown as D in Fig. 4.

$$D_{ij}(S,T) = \left| \frac{S_{ij} - T_{ij}}{S_{ij} + T_{ij}} \right|, \quad \text{where i} < \text{j} \qquad (4)$$

S_{ij} and T_{ij} are (i,j) elements in S and T. Since S_{ij} and T_{ij} are invariant features, D_{ij} also becomes an invariant and robust feature. For speaker-based clustering of World Englishes, we use D_{ij} as a feature in support vector regression.

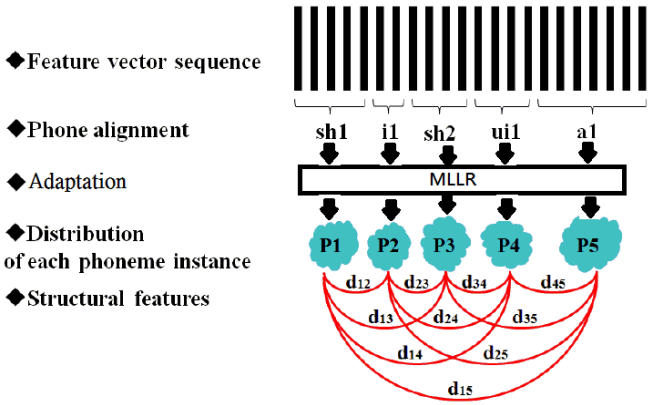

Fig.3 Extraction of structural features

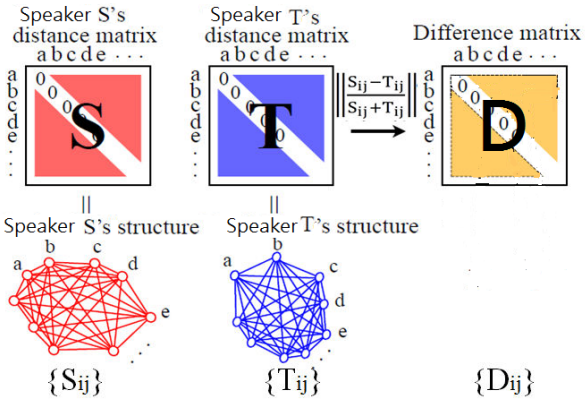

Fig.4 Inter-speaker structure difference

5. SVR to Predict Pronunciation Distances among Speakers

Using the reference distance between any two speakers and the upper triangle elements of difference matrix *D* between them, we trained a support vector regression (SVR). In this paper, LIBSVM [9] was adopted to train the SVR. Here, the SVR is expected to predict the reference distance using the upper triangle elements of difference matrix *D* as input, which are also called attribute values in LIBSVM. The reference inter-speaker distances were used as the target values. In this paper, the epsilon-SVR is used. The kernel type is radial basis function: *exp(-gamma * |x1-x2|^2)*.

We divided the elicitation paragraph into 9 sentences. Therefore 9 pronunciation structures were obtained, one for each sentence. From all of the 9 structures, a set of 2,804 phone distances were obtained for each speaker. An inter-speaker distance vector between two speakers, corresponding to the upper triangles of 9 difference matrixes, was also represented as a set of 2,804 values.

For performance evaluation, correlation between the reference distances and the predicted distances was used. We divided all the speaker pairs into 2 sets based on the reference distances and performed a 2-fold cross-validation (1 set was used to train SVR and the other set was used for testing). The correlation results of the first set and second set were 0.825 and 0.826, respectively. Fig. 5 shows the correlation results of these two sets.

For comparison, we also constructed a baseline system, which corresponds to an automated version of the inter-speaker distance calculation procedure described in section 3. The procedure is composed of two steps: 1) IPA manual transcription and 2) DTW-like algorithm for distance calculation. In the baseline system, the process of 1)

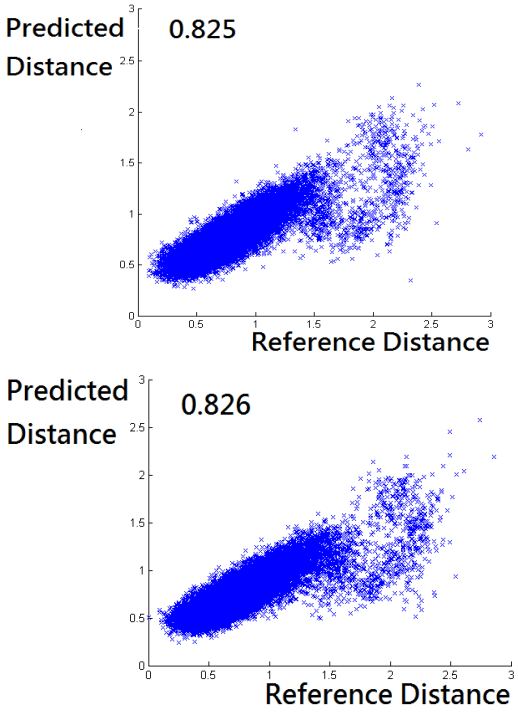

Fig.5 Correlation results of the two testing sets

is replaced with automatic recognition of phonemes in input utterances. Here, monophone HMMs trained using Wall Street Journal corpus were used. Since IPA transcription is based on phones and HMMs are trained based on phonemes, each IPA transcription has to be converted to its phoneme transcription. For conversion, the phonemes used in the CMU dictionary were adopted and each IPA symbol in the transcription was converted into its phonemic counterpart using a simple mapping table. Since this conversion may work as abstraction process, some detailed phonetic information will be lost in the conversion. However, the correlation between IPA-based speaker distances and phoneme-based speaker distances was found to be as high as 0.97. This means that the information loss is very minor and a perfect phoneme recognizer could predict inter-speaker distances very accurately.

The WSJ-based phoneme recognizer was used to recognize phonemes in wav files of the speech accent archive directly. Word-based network grammar was built for each sentence and used for recognition. Fig. 6 shows an example of word-based network grammar. In this figure, W_{ij} denotes the i-th word and the j-th possible pronunciation corresponding to the i-th word. Based on the recognition results, phoneme-level pronunciation error can be detected. In this paper, two kinds of grammar were used, which differ in whether the phonemic transcription of input utterance is included (closed) or not (open) in the grammar.

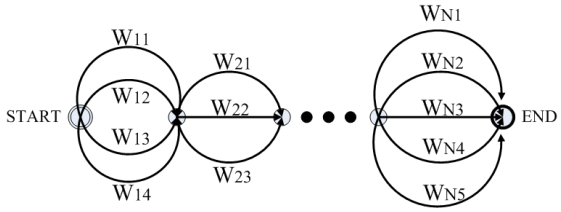

Fig.6 An example of word-based grammar

In the closed mode, the network was generated from all the 381 phonemic transcriptions.

The phone recognition accuracy of using the open word-based and the closed word-based grammar was 46.07% and 46.15%, respectively. The recognition results were used to estimate inter-speaker distances by comparing two phoneme sequences directly using the DTW-like algorithm and the penalty table introduced in the section 3. The correlation between the estimated inter-speaker distances using the open grammar and the reference IPA-based distances was 0.09 and that between the distances using the closed grammar and the IPA-based distances was also 0.09. These results mean our proposed method is much more robust than the conventional ones.

Based on the predicted inter-speaker distances by our proposed method, hierarchical speaker-based pronunciation clustering can be conducted. Since the clustering result of the 381 speakers is too complicated, we show here the result of selected speakers. We picked up Cantonese speakers in the archive, the number of which was found to be 7, and 7 American speakers were also selected. The clustering result of the 14 speakers using the predicted pronunciation distances is shown in Fig. 7.

In Fig. 7, "ca" and "en" denote Cantonese and native English (American) speakers, respectively. The attached numbers after "ca" and "en" are the speaker IDes tagged in the corpus. From this figure, we can say that the speakers are clustered into two big clusters mainly. One can be viewed as Cantonese sub-tree and the other as native English sub-tree. Only one Cantonese and one native English speaker were clustered into the contrary cluster. The clustering result shows that most speakers can be clustered correctly based on their accents.

6. Conclusions

With the ultimate aim of drawing the global map of World and individual Englishes, this paper investigated invariant pronunciation structure and SVR to predict inter-speaker pronunciation distances for new speaker pairs. The speech

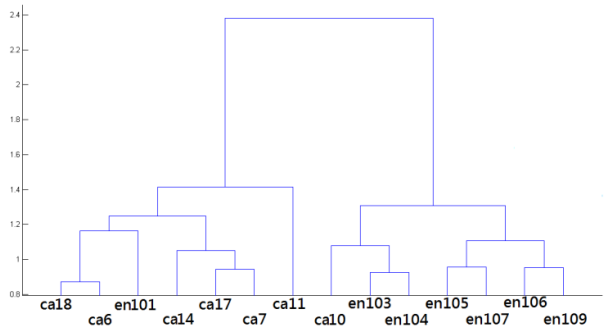

Fig.7 Speaker-based pronunciation clustering of 7 Cantonese speakers and 7 English speakers

accent archive, containing data from worldwide accented English speech, was used as training and testing samples. Evaluation experiments showed very promising results. The result achieved using our proposed method outperformed the result achieved using the conventional ones. In future work, we are planning to further define the list of penalties, which may be obtained by acoustic analysis of every phone pair spoken by a single speaker. Moreover, a more extensive collection of data is planned using smart phones and social network infrastructure such as crowdsourcing. Pedagogical application of the World and individual English map will also be considered in collaboration with language teachers.

References

[1] D. Crystal, *English as a global language*, Cambridge University Press, New York, 1995.

[2] J. Jenkins, *The phonology of English as an international language*, Oxford University Press, 2000

[3] M. Pinet *et al*. "Second-language experience and speech-in-noise recognition: the role of L2 experience in the talker-listener accent interaction", Proc. SLaTE, CD-ROM, 2010.

[4] S. H. Weinberger, *Speech Accent Archive*, George Mason University http://accent.gmu.edu.

[5] N. Minematsu, *et al*., *Speech structure and its application to robust speech processing*, Journal of New Generation Computing, 28, 3, pp. 299-319, 2010.

[6] X. Ma, *et al*. i, *Dialect-based speaker classification using speaker invariant dialect features,* in Proc. Int. Symposium on Chinese Spoken Language Processing, pp.171-176, 2010.

[7] The CMU pronunciation dictionary, http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[8] R. Jakobson *et al*., *The sound shape of language*, Mouton de Gruyter, 2002.

[9] C. Chang et al., LIBSVM: *a library for support vector machines*, 2001