INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Relationships between Perceptual Distinctiveness, Articulatory Complexity and Functional Load in Speech Communication

Yuqing Zhang[1], *Zhu Li*[1], *Bin Wu*[2], *Yanlu Xie*[1], *Binghuai Lin*[3], *Jinsong Zhang*[1]

[1]Beijing Language and Culture University, China
[2]Nara Institute of Science and Technology, Japan
[3]Smart Platform Product Department, Tencent Technology Co., Ltd, China

{yuqingelsa, lzblcu19}@gmail.com, wu.bin.vq9@is.naist.jp,
{xieyanlu, jinsong.zhang}@blcu.edu.cn, binghuailin@tencent.com

Abstract

Work on communicative efficiency has hypothesized that phonological contrasts signaling more meaning distinctions (i.e., of high functional load (FL)) tend to have the least articulatory complexity and the highest perceptual salience. However, only a few studies have examined the preference for perceptual distinctiveness based on the traditional measures of FL (e.g., the number of minimal pairs, the change in entropy of the lexicon), which are weak in modeling contexts of individual words. And little attention has been devoted to investigating the need to minimize effort. This study explores whether and how the communicative pressures to minimize the likelihood of confusion and minimize articulatory effort influence phonemic contrasts' functional contributions to speech communication. We used a revised definition of FL capable of modeling contextual information (i.e., the change in mutual information between phoneme sequences and spoken texts after the contrast in question is neutralized) and quantified information contributions of phonemic contrasts in English. The results indicated that FL of each phoneme pair increased significantly with its perceptual distinctiveness, and decreased significantly with articulatory complexity of the phoneme requiring less articulatory effort in the contrast. Altogether, these findings suggest that communicative pressures modulate the work a phonemic contrast does in distinguishing words.

Index Terms: communicative efficiency, functional load

1. Introduction

Communicative efficiency has been found to be one of the driving forces in shaping linguistic structure [1–6]. An efficient communicative system will allow for more information to be transmitted given the same amount of time. One well-known pattern expected from communicative efficiency is the avoidance of perceptually confusable linguistic units in human language, which enables the accurate word identification from the speech signal (see [1], for the communicative goals to maximize the distinctiveness of contrasts and to minimize articulatory effort). An example of such bias against perceptual confusion is the possible relationship between perceptual distinctiveness and FL. Studies have hypothesized that phonological contrasts of high FL tend to be highly perceptible [2, 7]. For example, in human language, the FL of the highly perceptible /b, t/ contrast would be relatively high, while the FL of the minimally perceptible /f, θ/ contrast would be comparatively low [8].

However, despite considerable attention devoted to speech perception and FL, the evidence concerning the effects of perceptual confusability on FL remains relatively scant. Gathering a sample of 49 languages from 25 language families, [2] examined place contrasts in stops and fricatives and provided preliminary supporting evidence that languages preferentially rely on perceptible phonemic contrasts to distinguish lexical minimal pairs. [9] investigated the relationship between auditory confusability and FL in written and spoken English, and suggested that due to pressures inherent in preventing communication failure in spoken speech, distinction exists in the structure of written and spoken lexicons. Contradictory evidence, however, has called into question the robustness of the relationship. For instance, [2] reported that in English, when the frequencies of the different sounds involved in the contrast were controlled, perceptual distinctiveness turned out to have no impact on the number of minimal pairs. Furthermore, though it has been acknowledged that the communicative pressure of minimizing articulatory effort plays an important role in shaping the structure of phonological systems [7], few studies have attested the effect of articulatory complexity on FL empirically, hence it remains an open question whether consonants involved in a sound contrast with high FL tend to have the least articulatory complexity.

Traditional measures of FL quantify the information contribution of a phoneme pair using the number of minimal pairs estimated based on word lemma types [10–12], or the change in word-level entropy of the lexicon upon merger of the phoneme pair [13–15]. However, these measures of FL can only represent a first approximation for a phonemic contrast's information contributions in speech communication, in that they may not take into account important properties of spoken word identification and have limited capacity to capture the information loss when the two minimally distinguished words are confused by listeners in actual language use. Research on spoken word recognition has indicated that different types of contexts (lexical, syntactic, semantic, and interpretative) influence the process of recognizing a spoken word [16]. Thus devising a FL measure effective in modeling contextual information of individual words is necessary in order to gain a closer approximation of the importance of a phoneme pair in speech communication.

In light of these limitations, we extend previous research and use a FL measure capable of modelling contextual information to ask whether and how the communicative pressures to keep words perceptually distinct and minimize articulatory effort affect information contributions of phoneme contrasts.

2. Methods

2.1. Text-Phoneme-Text transmission model

Speech communication is simulated using the Text-Phoneme-Text transmission model, as illustrated in Figure 1. F and F_α

http://dx.doi.org/10.21437/Interspeech.2021-721

denote the phonological transcriptions of the text W based on the original phoneme sets Φ, and the new phoneme sets Φ_α generated by merging the phonemic categories α. $\hat{W}$ and $\hat{W}_\alpha$ stand for the decoded word sequences from F and F_α respectively. As done in most speech recognition systems, the decoding process relies on word lattice scoring using the language model (LM) and the pronunciation lexicon with Φ or the lexicon with Φ_α.

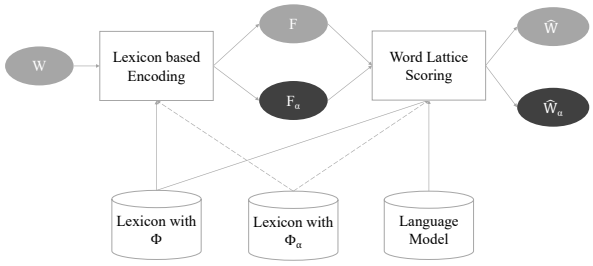

Figure 1: *The Text-Phoneme-Text transmission model.*

The conversion from W to F is analogous to speakers' encoding of communicable messages into sound sequences, which were received by listeners. F to $\hat{W}$ represents listeners' interpretation of phoneme sequences into meaningful words based on higher level knowledge. The path W-F_α-$\hat{W}_\alpha$ depicts mistransmission of the phonemic categories in α, which might result from channel noise interferences, misperception by the listener or lack of articulatory effort on the part of the speaker.

2.2. FL as change in mutual information

FL of a phoneme pair was calculated as change in mutual information of spoken texts (W) and phoneme sequences (F) induced by the merger of a phoneme pair α, as shown in formula 1. This algorithmic approach has been applied to analyzing the phonological system of Mandarin Chinese [17–19].

$$\mathrm{FL}(\alpha) = \frac{\mathrm{I}(\mathrm{W};\mathrm{F}) - \mathrm{I}\left(\mathrm{W};\mathrm{F}_\alpha\right)}{\mathrm{I}(\mathrm{W};\mathrm{F})} \tag{1}$$

The basic idea lies in that, upon merger of a phonemic contrast, the number of word sequences sharing the same phoneme transcription will increase, hence the mutual information between the spoken text and the phoneme transcription will decrease as compared to before. The mutual information loss reflects reduction of the amount of shared information due to the merger of a phoneme pair, and thus it can be utilized to quantify information contributions of phonemic contrasts.

According to the Shannon-McMillan-Breiman theorem [20], we can mathematically derive the formula 2, in which $W'_1, W'_2, \ldots, W'_m$ are all text sequences sharing the same transcription F. The probability of the text sequence $P\left(W'_i\right)$ can be efficiently computed by LMs.

$$\mathrm{I}(\mathrm{W};\mathrm{F}) = \lim_{\mathrm{n}\to\infty} -\frac{1}{\mathrm{n}} \log \sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{P}\left(\mathrm{W}'_\mathrm{i}\right) \tag{2}$$

We can use word hypothesis graph (WHG) to visualize our revised definition of FL, as shown in Figure 2. Intuitively, after merging a phoneme pair, there will be more text sequences within a WHG. Consequently, mutual information between the text W and its phonological transcription F will get smaller.

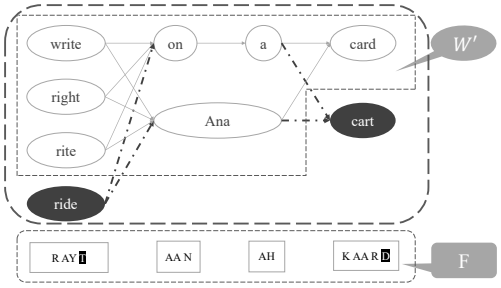

Figure 2: *The lattice of all the word sequences sharing the same phonological transcription F before and after /t, d/ are merged.*

2.3. Perceptual distictiveness

A confusion matrix captures the frequencies of phoneme identification errors, and therefore can be used to compute perceptual similarity/distinctiveness estimates. Among several available measures, we applied the phi-square statistic to confusion data from a phoneme identification experiment [21]. The phi-square statistic characterizes the perceptual distinctiveness of two phonemes x and y, derived from quantifying the similarity of the response distributions of two phonemes [22]. It is expressed mathematically as:

$$\Phi^2 = \sqrt{\frac{\sum \frac{(x_i - E(x_i))^2}{E(x_i)} + \sum \frac{(y_i - E(y_i))^2}{E(y_i)}}{N}} \tag{3}$$

N is the total number of responses, x_i and y_i equal the frequencies with which x and y are identified as category i. E(x_i) and E(y_i) represent the expected frequencies of responses for x_i and y_i if phonemes x and y are perceptually equivalent. The phi-square statistic reaches a value of zero when the distributions of phoneme identification responses are identical (i.e., two phonemes are maximally confusable), and reaches a value of one when the distributions have no overlap (i.e., two phonemes are perceptually distinct).

2.4. Articulatory complexity

Articulatory complexity values of English consonants were derived from [23]'s model reflecting the growth in motor control required to articulate increasingly complex consonants, as cited in [15]. In this model, it is formulated that plosives, nasals, and glides at the bilabial, alveolar, and glottal places of articulation have the least articulatory complexity in human language (e.g., /p, m, n, w, h/ in English). Values of articulatory complexity for English consonants range from 1 to 4, as summarized in Table 1.

Table 1: *Articulatory complexity of English consonants.*

Value	English consonants
1	m, p, w, h, n
2	k, b, g, j, f, d
3	t, r, l, ŋ
4	s, z, ʃ, ʒ, tʃ, dʒ, v, θ, ð

3. Experiment

3.1. Corpus processing

Corpora of subtitles offer a close approximation of daily informal conversations [24], and are thus suitable for addressing questions related to speech communication. For the present study, we randomly sampled 425,500 sentences (around 2,000,000 word tokens) from the "TV and Movies subtitles" (TVM90-19) sub-corpus of the Corpus of Contemporary American English [25] and converted them into ARPAbet sequences using the CMU dictionary[1]. Then we decoded ARPAbet sequences into all possible word sequences using the two lexicons. Finally we trained a trigram LM with the TVM90-19 corpus using KenLM [26] to score the lattice of all word sequences sharing the same phonological transcription, after which we got the FL value for each phoneme pair.

3.2. Statistical analysis and visualization

This study aims to examine the effect of perceptual distinctiveness and articulatory complexity on information contributions of phonemic contrasts. It is important to control the base frequencies of different sounds in order to draw inferences about the extent to which FL relies on the confusability of particular contrasts or the articulatory complexity of phonemes involved in the phoneme pair. Here, we computed the log-frequency of the more frequent sound (e.g., /f/ of the pair /f, θ/; coded as Log Frequency1) and the log-frequency of the less frequent sound (e.g., /θ/ of the pair /f, θ/; Log Frequency2) as control variables.

Multiple regression models were fitted to predict log-transformed FL values of 276 English consonant pairs from perceptual distictiveness, articulatory complexity of the phoneme requiring more articulatory effort (e.g., /ʃ/ of the pair /h, ʃ/; coded as Articulatory Complexity1), articulatory complexity of the phoneme requiring less articulatory effort (e.g., /h/ of the pair /h, ʃ/; Articulatory Complexity2), Log Frequency1 and Log Frequency2. Raw FL and frequency values were log-transformed as they exhibited positive skew and resulted in poor model fit. Contrasts with zero FL value (4 pairs) were removed from final analysis. All predictors were standardized. The final models had low collinearity [VIF's < 1.5] and visual inspections of residual plots revealed no obvious violations of the normality and constant variance assumptions.

Given a dissimilarity matrix with the distances between each pair of objects, multidimensional scaling (MDS) can yield spatial representation of each object while preserving the original between-object distances as much as possible. Using this visualization technique, we analyzed the distribution of phonemes based on FL and compared that to the perceptual space. Since FL stands for the work a phonemic contrast α does in differentiating words, the pairwise distance was defined as:

$$d(\alpha) = 1 - FL(\alpha) \tag{4}$$

such that phoneme pairs with high FL would be put close together (Figure 4). The perceptual space was constructed using Phi-square statistic as pairwise perceptual distance (Figure 5).

4. Results and Discussion

4.1. Effect of phoneme frequency on FL

As shown in Table 2, pairwise FL increases significantly with the log-frequency of the less frequent sound as well as the log-frequency of the more frequent sound, meaning that contrasts' functional contribution to speech communication is dependent on the frequencies of the different sounds involved in that contrast. Of all predictors considered in the model, the log-frequency of the less frequent sound has the strongest relationship with FL ($\beta = 0.68$, t = 16.21, $p < 0.0001$). As shown in Figure 4, phonemes /ʒ/, /ŋ/ and /θ/ are far away from other phonemes, suggesting that the phoneme pairs involving them are ranked bottom on the FL metrics, which is primarily because these three phonemes are among the least frequent sound in English. In addition, phonemes /h/, /ʃ/ and /w/ are grouped closely together, which is largely due to the fact that the pronouns 'he' 'she' and 'we' are highly frequent. These results echo prior findings in other languages. [18] reported that phonemes /l/ and /d/ had the highest FL in Mandarin, which might be partially accounted for by the high frequency of particles 'le' and 'de' in communication. And the three-way retroflex contrast of voiceless affricates and fricatives /tʂ, tʂʰ, ʂ/ (zh, ch, sh in Pinyin Romanization) had higher pairwise FL value than the dental series /ts, tsʰ, s/ (z, c, s in Pinyin Romanization), which in part results from the high frequency of the retroflex sounds.

Table 2: *Standardized regression coefficients (predicting LogFL of phoneme pairs) for the multiple regression model.*

	Coefficient	SE	t	p
Log Frequency1	0.185	0.039	4.754	< 0.001 ***
Log Frequency2	0.680	0.042	16.210	< 0.001 ***
Perceptual Distinctiveness	0.126	0.037	3.452	< 0.001 ***
Articulatory Complexity1	-0.035	0.041	-0.848	0.397
Articulatory Complexity2	-0.113	0.040	-2.830	0.005 **

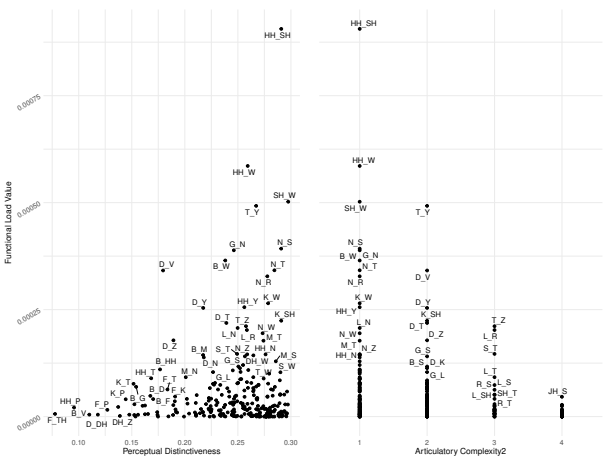

Figure 3: L_{eft}: *Raw FL values for 276 English consonant pairs (y-axis) plotted against perceptual distinctiveness values inferred by applying Phi-square statistic to [21]'s data (x-axis; numbers closer to zero indicate higher confusability).* R_{ight}: *Raw FL values plotted against articulatory complexity values of the phonemes requiring less articulatory effort in sound pairs.*

4.2. Effect of perceptual distinctiveness on FL

In line with prior studies [9], our results demonstrated the predicted effect of perceptual distinctiveness on FL, i.e., the FL value for each contrast increases significantly with its perceptual distinctiveness ($\beta = 0.13, t = 3.45, p < 0.001$). As

[1] https://github.com/Alexir/CMUdict/blob/master/cmudict-0.7b

shown in Figures 3 and 4, the phoneme pair /h, ʃ/ has the highest FL, which partly results from the perceptual similarity between them being much weaker than other pairs, as these two phonemes are far away from each other in the perceptual space (Figure 5). The /f, θ/ pair has the highest probability of being confused, and correspondingly their FL value is minimal. In other words, minimal pairs like fink:think are less likely, while minimal pairs like heat:sheet should be more likely to occur. In addition, we found that consonants with the same place of articulation but different manners tend to be grouped closer (e.g., /l, r, n, t/, /b, m/ in Figure 4), indicating that phonemes tend to have higher pairwise FL within these groups. This observation supports the hypothesis that an efficient communicative system is biased for perceptual distinctiveness, as [8, 27] have suggested that most phonemes in English at the same place of articulation with different manners tend to have long perceptual distance.

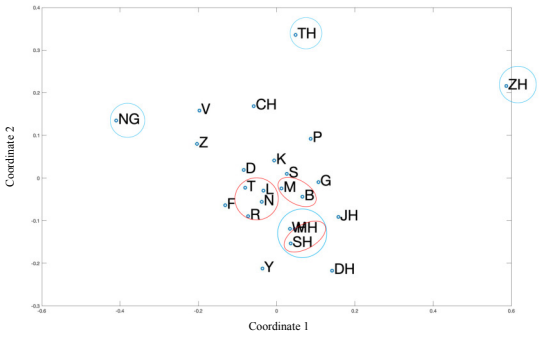

Figure 4: *MDS visualization of English consonants based on FL. Phoneme pairs of high FL are put close together.*

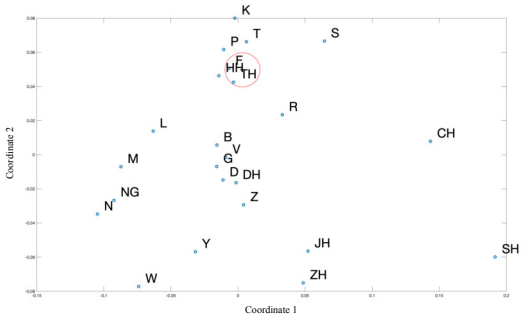

Figure 5: *MDS visualization of English consonants based on* Φ^2*. Phoneme pairs of high perceptual similarity are closer.*

4.3. Effect of articulatory complexity on FL

Our results supported the predicted effect of articulatory complexity on FL, i.e., FL for each contrast decreases significantly with articulatory complexity of the phonemes involved in the contrast. For English consonants, FL was negatively associated with articulatory complexity of the phoneme requiring less articulatory effort ($\beta = -0.11, t = -2.83, p < 0.01$). It should be noted that articulatory complexity of the phoneme requiring more articulatory effort has no significant effect on information contribution of phoneme contrasts ($\beta = -0.03, t = -0.85, p > 0.1$). This phenomenon cannot be explained in terms of communicative efficiency. We speculate that if one contrast includes a phoneme requiring the least articulatory effort, FL of this contrast would be high regardless of whether the other phoneme has high or low articulatory complexity.

Table 3 shows the five highest-ranked segments as well as the five highest-ranked contrasts. Here, the FL of a phoneme φ was computed by summing all pairwise FL over contrasts involving the phoneme φ and then applying the normalizing factor 1/2. Consistent with our prediction, the top three consonants /w, n, h/ have the least articulatory complexity (i.e., their articulatory complexity values equal to 1). And out of the top ten consonant pairs with higher FL, only two pairs don't involve the consonants requiring the least articulatory effort.

Table 3: *5 consonants and consonant pairs with the highest FL.*

Rank	Consonant	FL	Consonant pair	FL
1	w	1.44	h-ʃ	0.91
2	n	1.42	h-w	0.59
3	h	1.37	ʃ-w	0.50
4	t	1.14	t-j	0.49
5	ʃ	1.04	n-s	0.39

Note: values should be multiplied by 0.001.

5. Conclusions

In a crosslinguistic corpus-based study of functional load and the organization of phonological systems, [28] noted that phonological properties may emerge from a set of non-linguistic (cognitive, motor, perceptual, communicative) abilities (originally proposed in [29]). Inspired by this account, the present study set out to explore the link between perceptual distinctiveness, articulatory complexity and FL in speech communication, and demonstrated how communicative pressures to minimize the likelihood of perceptual confusion and minimize articulatory effort influence phonemic contrasts' functional contributions to speech communication while the effect of phoneme frequencies is controlled. Information contributions of phoneme pairs were quantified using our revised definition of functional load effective in modeling contextual information of spoken words. Results indicated that consistent with the predictions derived from communicative efficiency, phonological contrasts of higher perceptual distinguishability or phonemes of less articulatory complexity do more work in identifying distinct words.

Future work will perform similar analyses in other languages, to see whether this is a universal pattern across all language users. Other directions include utilizing state-of-the-art LMs to estimate probabilities for a sentence (e.g., masked LMs [30, 31] and autoregressive LMs [32]), so as to obtain a more precise estimate of FL for each phonemic contrast.

6. Acknowledgements

This study was supported by Advanced Innovation Center for Language Resource and Intelligence (KYR17005), National Social Science Foundation of China (18BYY124), Wutong Innovation Platform of Beijing Language and Culture University (19PT04), the Science Foundation and Special Program for Key Basic Research fund of Beijing Language and Culture University (the Fundamental Research Funds for the Central Universities) (21YJ040004, 21YCX180). The first two authors provide equal contribution. Jinsong Zhang is the corresponding author.

7. References

[1] E. S. Flemming, *Auditory representations in phonology*. Routledge, 2013.

[2] P. N. H. M. Graff, "Communicative efficiency in the lexicon," Ph.D. dissertation, Massachusetts Institute of Technology, 2012.

[3] G. K. Zipf, *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books, 2016.

[4] E. Gibson, R. Futrell, S. P. Piantadosi, I. Dautriche, K. Mahowald, L. Bergen, and R. Levy, "How efficiency shapes human language," *Trends in cognitive sciences*, vol. 23, no. 5, pp. 389–407, 2019.

[5] S. T. Piantadosi, H. Tily, and E. Gibson, "Word lengths are optimized for efficient communication," *Proceedings of the National Academy of Sciences*, vol. 108, no. 9, pp. 3526–3529, 2011.

[6] M. Hahn, D. Jurafsky, and R. Futrell, "Universals of word order reflect optimization of grammars for efficient communication," *Proceedings of the National Academy of Sciences*, vol. 117, no. 5, pp. 2347–2353, 2020.

[7] E. Flemming, "Contrast and perceptual distinctiveness," *Phonetically based phonology*, vol. 232, p. 276, 2004.

[8] G. A. Miller and P. E. Nicely, "An analysis of perceptual confusions among some English consonants," *The Journal of the Acoustical Society of America*, vol. 27, no. 2, pp. 338–352, 1955.

[9] S. Kang and C. Cohen, "Relationships between functional load and auditory confusability under different speech environments." in *INTERSPEECH*, 2016, pp. 2821–2825.

[10] D. Ingram and I. David, *First language acquisition: Method, description and explanation*. Cambridge university press, 1989.

[11] A. Wedel, S. Jackson, and A. Kaplan, "Functional load and the lexicon: Evidence that syntactic category and frequency relationships in minimal lemma pairs predict the loss of phoneme contrasts in language change," *Language and speech*, vol. 56, no. 3, pp. 395–417, 2013.

[12] A. Wedel, A. Kaplan, and S. Jackson, "High functional load inhibits phonological contrast loss: A corpus study," *Cognition*, vol. 128, no. 2, pp. 179–186, 2013.

[13] D. Surendran and P. Niyogi, "Measuring the functional load of phonological contrasts," *arXiv preprint cs/0311036*, 2003.

[14] ——, "Quantifying the functional load of phonemic oppositions, distinctive features, and suprasegmentals," *Amsterdam studies in the theory and history of linguistic science series 4*, vol. 279, p. 43, 2006.

[15] S. F. Stokes and D. Surendran, "Articulatory complexity, ambient frequency, and functional load as predictors of consonant development in children," *Journal of Speech, Language, and Hearing Research*, pp. 577–591, 2005.

[16] U. H. Frauenfelder and L. K. Tyler, "The process of spoken word recognition: An introduction," *Cognition*, vol. 25, no. 1-2, pp. 1–20, 1987.

[17] J. Zhang, W. Li, Y. Hou, W. Cao, and Z. Xiong, "A study on functional loads of phonetic contrasts under context based on mutual information of Chinese text and phonemes," in *2010 7th International Symposium on Chinese Spoken Language Processing*. IEEE, 2010, pp. 194–198.

[18] B. Wu, J. Zhang, and Y. Xie, "A clustering analysis of Chinese consonants based on functional load," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*. IEEE, 2014, pp. 1–4.

[19] Y. Chen, Y. Xie, and J. Zhang, "A comparison study of information contributions of phonemic contrasts in Mandarin," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 1579–1582.

[20] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[21] A. Cutler, A. Weber, R. Smits, and N. Cooper, "Patterns of English phoneme confusions by native and non-native listeners," *The Journal of the Acoustical Society of America*, vol. 116, no. 6, pp. 3668–3678, 2004.

[22] P. Iverson, L. E. Bernstein, and E. T. Auer Jr, "Modeling the interaction of phonemic intelligibility and lexical structure in audiovisual word recognition," *Speech Communication*, vol. 26, no. 1-2, pp. 45–63, 1998.

[23] R. D. Kent, "The biology of phonological development," *Phonological development: Models, research, implications*, vol. 65, p. 90, 1992.

[24] N. Levshina, "Online film subtitles as a corpus: An n-gram approach," *Corpora*, vol. 12, no. 3, pp. 311–338, 2017.

[25] M. Davies, "The Corpus of Contemporary American English as the first reliable monitor corpus of English," *Literary and linguistic computing*, vol. 25, no. 4, pp. 447–464, 2010.

[26] K. Heafield, "Kenlm: Faster and smaller language model queries," in *Proceedings of the sixth workshop on statistical machine translation*, 2011, pp. 187–197.

[27] J. Zhang, S. Lu, and S. Qi, "A cluster-analysis of the perceptual features of Chinese speech sounds," *Journal of Chinese linguistics*, vol. 10, no. 2, pp. 190–206, 1982.

[28] Y. M. Oh, C. Coupé, E. Marsico, and F. Pellegrino, "Bridging phonological system and lexicon: Insights from a corpus study of functional load," *Journal of phonetics*, vol. 53, pp. 153–176, 2015.

[29] C. Moulin-Frier, J. Diard, J.-L. Schwartz, and P. Bessière, "Cosmo ("Communicating about objects using sensory–motor operations"): A bayesian modeling framework for studying speech communication and the emergence of phonological systems," *Journal of Phonetics*, vol. 53, pp. 5–41, 2015.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[31] J. Salazar, D. Liang, T. Q. Nguyen, and K. Kirchhoff, "Masked language model scoring," *arXiv preprint arXiv:1910.14659*, 2019.

[32] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.