INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Quantifying expectation modulation in human speech processing

M. Bentum[1], *L. ten Bosch*[1,2], *A van den Bosch*[1,3], *M. Ernestus*[1,2]

[1]Center for Language Studies, Radboud University, Nijmegen, the Netherlands
[2]Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands
[3]KNAW Meertens Institute, Amsterdam, the Netherlands
{m.bentum, l.tenbosch, a.vandenbosch, m.ernestus} @let.ru.nl

Abstract

The mismatch between top-down predicted and bottom-up perceptual input is an important mechanism of perception according to the predictive coding framework (Friston, [1]). In this paper we develop and validate a new information-theoretic measure that quantifies the mismatch between expected and observed auditory input during speech processing. We argue that such a mismatch measure is useful for the study of speech processing. To compute the mismatch measure, we use naturalistic speech materials containing approximately 50,000 word tokens. For each word token we first estimate the prior word probability distribution with the aid of statistical language modelling, and next use automatic speech recognition to update this word probability distribution based on the unfolding speech signal. We validate the mismatch measure with multiple analyses, and show that the auditory-based update improves the probability of the correct word and lowers the uncertainty of the word probability distribution. Based on these results, we argue that it is possible to explicitly estimate the mismatch between predicted and perceived speech input with the cross entropy between word expectations computed before and after an auditory update.

Index Terms: speech perception, predictive coding, statistical language modelling, automatic speech recognition

1. Introduction

Listeners are able to extract words from speech input in a wide range of (adverse) listening conditions. The difficulty of this task is attested by the many decades of research aimed at creating artificial systems with similar performance. The details of the cognitive processes underlying human speech processing are still contentious. A long-standing debate revolves around the importance and timing of top-down versus bottom-up influence for word recognition during speech comprehension [2,3]. Certain autonomous models (e.g. Shortlist A & B [4,5]) claim that early speech processing is exclusively bottom-up, and top-down influence is only exerted at the lexical phase of word recognition. Other interactive models (e.g. Trace [6]) allow for a certain degree of top-down influence, congruent with the predictive coding framework [1].

The predictive coding framework [1] assumes that perception entails anticipation based on a generative model, whereby cognitively higher levels generate predictions about upcoming (low-level) perceptual input. The mismatch between the prediction and the actual input provides an error signal, which informs to what extent the hypotheses generated by the generative model need to be adapted. If we think about human speech processing in this framework, we need a model to assign a probability to upcoming words, given the preceding words, and a mechanism to quantify the mismatch between bottom-up observations and top-down expectations. The first part, the probability of upcoming words, can be estimated according to Equation 1, which lies at the basis of a statistical language model (SLM), whereby *P* denotes the conditional probability of word *Wi* given a sequence of *n* preceding words:

$$P(W_i) = P(W_i|W_{i-n}, \ldots, W_{i-1}) \quad (1)$$

Several studies (e.g. [7,8,9]) have successfully used statistical language modelling to study human language processing. They employed an SLM to compute word probabilities from a text corpus and show that listeners and readers are indeed sensitive to the probability of a word given the preceding words. These results suggest that listeners anticipate likely upcoming words. The predictive coding framework makes an additional prediction, namely that human listeners generate low-level auditory expectations based on the anticipated words.

This paper addresses the estimation of the mismatch (i.e. the error signal) between the expected word form and the observed word form as it comes in as speech input. To estimate this error signal, we make use of the concept of a word probability distribution (WPD), consisting of a list of words, whereby each word is assigned a probability. We compute two types of WPD. The *prior* WPD is based on the top-down expectation at word onset without any auditory input. In this WPD, each word is assigned a probability given the preceding words as estimated by an SLM. A *post* WPD is based on the prior WPD in combination with the bottom-up acoustic evidence received so far: i.e. the word probabilities are updated according to the unfolding auditory information.

We analyze the auditory input with statistical paradigms developed in the automatic speech recognition (ASR) domain, to generate a probability distribution on a large set of phone sequences that could all potentially match a possible word start. These phone sequence probabilities are used to update word probabilities matching these phone sequences, resulting in a post WPD. The error signal can then be defined as the cross entropy between the prior and post WPD, which captures the mismatch between the high-level expectation (word probabilities) and the sensory input (a spoken word). The cross entropy between prior and post WPD can be computed with Equation 2, whereby *H* denotes cross entropy, *p* the prior WPD, *q* the post WPD and *X* the WPD word list.

$$H(p,q) = -\sum_{x \in X} p(x) \log q(x) \quad (2)$$

To summarize, the prior WPD captures high-level expectations (based on preceding words). The post WPD differs only from the prior WPD in the added auditory information. We therefore

http://dx.doi.org/10.21437/Interspeech.2019-2685

propose that the cross entropy between prior and post WPD quantifies the mismatch between high-level expectations and auditory input.

To validate the computation of the mismatch measure, we test if the auditory update improves the post WPD in relation to the prior WPD. We expect the auditory update to decrease the entropy of the post WPD and increase the probability of the correct word. In addition, we test whether these measures improve with more auditory input. Since our goal is to compute a mismatch measure which is relevant for human speech processing, we also test the optimal amount of auditory materials for cross entropy computation. In the following sections, we will describe the language materials and methods used to compute both the prior and post WPDs and the subsequent analyses. After these sections, results are presented, followed by a discussion and a future outlook.

2. Method

2.1. Materials

We used materials from three corpora, namely, the Spoken Dutch Corpus [10], IFADV [11] and NLCOW14, henceforth COW [12,13]. The first two corpora consist of audio recordings and transcriptions of spoken Dutch materials. The COW corpus consists of 4,7 billion words of web-crawled Dutch text.

We pre-processed the COW corpus by excluding all non-Dutch sentences, removing sentences with three or more repeating words or characters, or characters that are not used in standard Dutch orthography. We replaced characters with diacritics to the equivalent characters without diacritics. Furthermore, we mapped all numbers, websites and tagged words (e.g. @tag@) to special word codes. We removed all punctuation, except for commas. We normalized all apostrophe words to a standard spelling (e.g. *'t* becomes *het*, 'the'). The Spoken Dutch Corpus and IFADV were already appropriately tokenized (see [14]); we only applied the apostrophe normalization and diacritic removal to these texts.

For our experiments we extracted a subset of the Spoken Dutch Corpus and the IFADV containing 50,277 word tokens (see Table 1). This subset, henceforth called Speech Corpus, consists of annotated speech from different speech registers (i.e. spontaneous dialogues, news broadcasts, and read aloud stories). The selection criteria for our materials were based on a different experiment. The differences in speech styles reflected in our materials will not be important in the current study.

Table 1: *Overview of the materials in the Speech Corpus.*

speech style	word tokens (word types)	average word duration (ms)
spontaneous dialogues	21,718 (2,435)	206
read-aloud stories	13,209 (2,349)	256
news broadcast	15,350 (3,526)	289
total	50,277 (5,866)	245

2.2. Procedure

For each word in the Speech Corpus we created two types of word probability distributions (WPD), one prior and one post auditory information integration (see Figure 1). We will explain how we created these WPDs for a given word (henceforth 'target word') in the Speech Corpus. To create the prior WPD, we used an SLM and a lexicon (i.e. the set of words in the WPD). We trained a 4th order Markov SLM on the Dutch COW corpus by using SRILM [15] with Kneser-Ney discounting for smoothing [16]. For the lexicon we selected approximately 200,000 Dutch phonemically transcribed words that are in the top .9 cumulative probability of the word unigrams of the SLM. We estimated the probability of each word in this lexicon based on the words preceding the target word in the Speech Corpus.

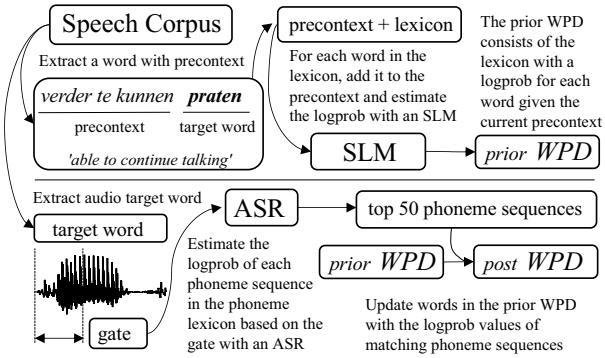

Figure 1: *Diagram of prior and post WPD construction.*

Post WPD construction was done in multiple steps. In the first step, we used the forced aligned phonemic transcriptions (present for all materials in the Speech Corpus) to determine the word onset of the target word in the audio materials and defined 28 gates of different durations (110, 130, …, 650 ms), starting from word onset. In step 2, each gate was used to create a post WPD, resulting in 28 post WPDs per target word. Figure 1, bottom part, shows post WPD construction for one gate.

To create the post WPDs, we used KALDI [17] to estimate a phonemic probability distribution for the gated speech input. We did this by first creating a 'Phoneme Lexicon' consisting of all lexically licensed phoneme sequences up to length 8, approximately 400,000 entries. For example, the word *universiteit* 'university' with phonemic representation /y n i v ɛ r s i t ɛi t/ yields the eight cohort forms /y/, /y n/, ..., /y n i v ɛ r s i/ to be included into the Phoneme Lexicon. This Phoneme Lexicon, in combination with a flat language model (i.e. each phoneme sequence has an equal prior probability), was used in the KALDI decoding of the gated speech chunks. For each gate, this decoding leads to a weighted phone lattice. The 500 best paths through this lattice were chosen as a decoding result. This step resulted in scaled logprob scores for each of the 500 phoneme sequences.

The scaled logprob scores were 'descaled' to a genuine probability distribution. The descaling factor determines the decay of the phoneme string probabilities (i.e. the probability difference between the winning hypothesis, runner-up, etc.). This descaling factor was estimated by investigating the entropy of the phonemic probability distribution for different gate durations. We assume that the entropy of the phoneme probability distribution should decrease for increasing gate lengths, because more acoustic material should yield a better identification and thereby a sharper distribution of the phoneme sequence probabilities. We therefore chose the factor which resulted in the highest entropy decrease across gates to descale the logprobs.

After descaling the logprobs, we inspected the phoneme *n*-best lists for multiple words from the Speech Corpus to determine a useful value of *n*. The top-50 appeared to be a sufficient threshold to exclude implausible phoneme sequence strings.

The logprobs of the top-50 phoneme strings were used to update the prior WPD to the post WPD. However, directly adding logprob values has (for our purposes) an unfortunate effect of generating the biggest difference in unlikely candidates. Since we truncated our *n*-best phoneme sequence set, this would result in a bad update. We therefore *shifted* the logprobs by adding the absolute value of the logprob of phoneme sequence 51 (from the *n*-best list) to the top-50 phoneme sequences. The most likely phoneme sequence now causes the biggest shift in the post WPD and normalization of this distribution ensures that unlikely words are shifted downwards appropriately.

To perform the auditory update, we matched each of the top-50 candidate phoneme sequences to all words in the lexicon (i.e. the set of words in the WPD). For example, the word *kat* 'cat', represented in the Dutch lexicon as 'kat, k ɑ t' would match with the phoneme sequences /k/, /k ɑ/, /k ɑ t/ and mismatch with /ɑ/, /ɑ t/ or /k ɑ t s/. We computed the word probabilities of the post WPD by adding the shifted logprob values of phoneme sequences to the logprob values of matching words in prior WPD.

2.3. Analysis

We performed two analyses to validate our approach and one to investigate the amount of auditory materials needed for the best cross entropy computation. For Analysis 1, we tested whether the auditory update from prior to post WPD lowered the surprisal of the correct word, which tests whether the auditory update assigns more probability to the correct word. Furthermore, we test whether the entropy of the post WPD was lower compared to the prior WPD, indicating that there is less uncertainty in the post WPD, which is expected if the auditory update functions correctly.

To make the comparison between prior and post WPD, we conducted two tests to check that both surprisal and entropy decrease after the auditory update. The first test was a conservative test that compares surprisal of the correct word and entropy of the prior WPD to the highest (i.e. worst) surprisal and entropy values of the set of 28 post WPDs for a given word. The less conservative test compared the surprisal of the correct word and entropy of the prior WPD to the mean of the surprisal and entropy over the same set of post WPDs. In both cases (conservative and less conservative), the post WPD surprisal and entropy values are compared with the corresponding prior WPD values.

For Analysis 2, we tested whether the surprisal value of the correct word and the entropy of the post WPDs decreased with increasing gate duration. We tested this by first computing the difference in surprisal of the correct word between prior and post WPD for each gate. Longer gates should improve the post WPD more, because a longer gate provides more information about the upcoming word. Of course, this only holds if the gate is shorter than the word, because otherwise information of following words is also incorporated in the auditory update. We therefore excluded all cases where the word was shorter than the gate.

Finally, Analysis 3 investigated which gate should be used for the cross entropy computation. We want to use the cross entropy to predict human speech processing cost and therefore we tested which gate duration performs the best update for all words (including words shorter than a given gate). This analysis reflects the situation for a human listener, who does not know the duration of upcoming words. For this analysis, we computed the difference in surprisal for the correct word between prior and post WPD for all words and gate durations.

3. Results

We used R [18] for all analyses. For Analysis 1, we compared the surprisal of the correct word between the prior and post WPD with a simple linear regression model. The regression model was fitted on 80% of the data and tested on 20% unseen data. Based on the results of the unseen data, we computed the R^2_{CV} (cross validated). Similar R^2 and R^2_{CV} values indicate that the model generalizes well to unseen data and was not overfitted to the current sample. We created separate models for the conservative (i.e. worst) and less conservative (average) test, as detailed in Section 2.2. We used the same approach to compare the entropy between prior and post WPD. As expected, both surprisal and entropy decrease (i.e. have negative betas) after the auditory update, as can be seen in Table 2. This appears both from the conservative and less conservative test.

Table 2: *Simple linear regression models for surprisal and entropy comparison between prior and post WPD.*

	R^2 (R^2_{CV})	B	SE B	P
worst surprisal	0.02 (0.02)			< .001
update*		-0.49	0.01	< .001
avg.† surprisal	0.64 (0.64)			< .001
update*		-2.80	0.01	< .001
worst entropy	0.14 (0.14)			< .001
update*		-1.85	0.02	< .001
avg.† entropy	0.80 (0.80)			< .001
update*		-6.30	0.01	< .001

*Difference between prior and post WPD, †average

For Analysis 2 we tested whether the surprisal of the correct word of the post WPD improves with increasing gate length. We fitted a linear regression model on the difference in surprisal between prior and post WPD for each gate length. We modelled the relationship between surprisal difference and gate length with a 7th order polynomial, to capture possible non-linear relationships and established the order of the polynomial with model comparison by selecting the highest uneven order that still improved the model. We used the same approach to test entropy difference in relation to gate length; for this model we used an 11th order polynomial on gate duration. Both the surprisal and the entropy model were fitted on 80% of the data. Again, we used the remaining 20% unseen data to compute the R^2_{cv}, to test whether the model generalizes well.

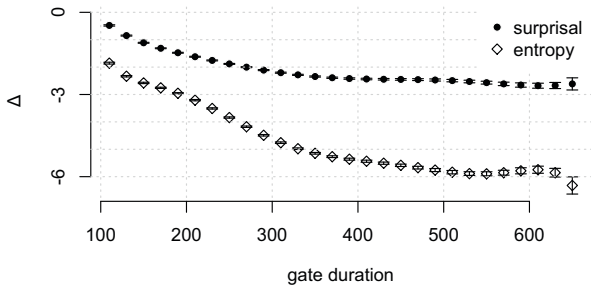

Figure 2: *Predicted difference in surprisal and entropy as a function of gate duration, with 99% confidence intervals.*

We do not report the betas for all polynomials in the surprisal and entropy model, because they are hard to interpret. Instead we visualized the results of both models in Figure 2. The surprisal of the correct word shows a clear negative trend with increasing gate length (R^2 = 0.13, R^2_{CV} = 0.14, p < .001). Similarly, the entropy of the post WPD also shows a clear

negative trend with increasing gate length ($R^2 = 0.22$, $R^2_{CV} = 0.22$, $p < .001$). The negative trend for surprisal means that with increasing gate length the probability of the correct word increases (if only words longer than the gate duration are considered). The negative trend for entropy means that the amount of uncertainty in the post WPD keeps decreasing when more relevant acoustic information becomes available.

Finally, we investigated which gate duration most improved the surprisal of the correct word for all words. We fitted a linear regression model on 80% of the data to predict the difference in surprisal by gate duration with a 7th order polynomial, $R^2 = 0.13$, $R^2_{CV} = 0.13$, $p < .001$. In Table 3 we report the top 3 gate durations that most improved the surprisal of the correct word after the auditory update. In addition, we fitted the same regression model on a randomly selected subset (10% of the data) a 1000 times. For each model we ranked the predicted surprisal difference, with rank 1 for best performance. Table 3 shows that the auditory update of 190 milliseconds resulted in the largest reduction in the surprisal of the correct word.

Table 3: *Predicted surprisal difference and number of times a gate duration (milliseconds) showed best improvement in surprisal between prior and post WPD.*

gate	predicted	99% CI	# rank 1
170	-0.978	-1.004, -0.952	285
190	**-0.982**	**-1.006, -0.958**	**715**
210	-0.944	-0.968, 0.920	0

4. Discussion

The goal of this study was to quantify a mismatch measure between high-level expectations and low-level input in speech perception. We created two types of word probability distributions (WPD), one prior and one post auditory update. The prior WPD is completely based on preceding words and represents the high-level expectations. The post WPD is an update of the prior WPD integrating auditory information. We hypothesized that the difference between prior and post WPD captures the mismatch between expectations and speech input and could be quantified by cross entropy.

To validate the mismatch measure, we investigated whether the auditory update performed as expected. In Analysis 1, we showed a decrease in both the surprisal of the correct word and the entropy of the post WPD, in line with our expectations. Furthermore, we showed in Analysis 2 that the surprisal and entropy further decrease with increasing gate length (only considering words that are longer than the gate duration). This was also expected; longer gate durations provide more information for the auditory update and should therefore improve update results.

The results show that the difference between the prior and post WPD reflects auditory information, which improved the probability of the correct word and lowered the uncertainty (entropy) of the post WPD. Prior and post WPD differ in word probabilities based on the extra information that the auditory input provides. We therefore argue that the cross entropy between both distributions captures the mismatch between expected and observed auditory input.

After validating our results, we investigated the amount of auditory materials needed to compute the mismatch measure in Analysis 3. For this analysis we included all words, because this more closely resembles the situation of a human listener (who does not know how long the next word will be). We compared surprisal improvement of the correct word between different gate durations and found that a gate of 190 milliseconds performed best. We also confirmed this with smaller subsets of the data, suggesting that this result generalizes to unseen data.

The mismatch measure we developed can be usefully applied in language research and could inform the discussion about autonomous versus interactive language processing. Norris et al. [19], arguing for the autonomous word recognition models, discusses the evidence pertaining predictive coding and suggests that more evidence is needed to provide insight for the role of predictive coding in language processing. The mismatch measure can elucidate whether cognitively high-level anticipations are relevant during the processing of low-level incoming speech sounds in human listeners, which, if found, would provide evidence against a strong autonomous bottom-up-only mechanism for speech perception (see below for a possible experiment).

A further question concerning the role of prediction in language processing is to what extent listeners predict speech input in regular non-experimental situations. Huettig [20] notes that most evidence for prediction in language processing comes from experiments that only investigated the extremes of predictability, comparing, for example, highly predictable words with unpredictable words. Recent studies (e.g. [7,8,9]) using information-theoretic measures, such as word surprisal and entropy to predict processing costs during language processing, investigate the whole spectrum of predictability. These studies show that human listeners and readers are sensitive to these information-theoretic measures across the whole predictability spectrum. Similarly, the mismatch measure we developed quantifies the whole range of mismatch between high-level expectations and low-level input. This will allow us to investigate the importance of predictive coding in regular speech processing.

A key test of the mismatch measure is to analyze its relation to data from human listeners. For example, in an experiment using electroencephalography (EEG) it has been shown that listeners are sensitive to violations of expected auditory forms [21,22]; this effect is referred to as the phonological mismatch negativity (PMN). We hypothesize that our measure should predict the amplitude of the PMN, whereby higher cross entropy between prior and post WPD would result in a more negative deflection of the EEG-signal.

5. Conclusions

The predictive coding framework proposes that the mismatch between cognitively high-level expectations and low-level perceptual input is an important mechanism in perception. We showed that we can quantify this mismatch for speech perception with the aid of statistical language modelling and an automatic speech recognition system. We used naturalistic speech recordings, containing approximately 50,000 words, to compute the mismatch measure. This opens up the possibility of investigating the importance of predictive coding during normal speech processing. We propose that the mismatch measure could be used to predict processing measures of listeners during speech perception. The results can inform the discussion about autonomous versus interactive models of speech perception.

6. Acknowledgements

We would like to thank Lou Boves for helpful discussions.

7. References

[1] K. Friston, "A theory of cortical responses," *Philosophical transactions of the Royal Society B: Biological sciences,* vol. 360, no. 1456, pp.815-836, 2005.

[2] J. S. Magnuson, D. Mirman, S. Luthra, T. Strauss, and H. D. Harris, "Interaction in spoken word recognition models: Feedback helps," *Frontiers in psychology*, vol. 9, no. 369, 2018.

[3] D. Norris, J. M. McQueen, and A. Cutler, "Commentary on 'Interaction in spoken word recognition models: Feedback helps'," *Frontiers in psychology,* vol. 9, no. 1568, 2018.

[4] D. Norris, "Shortlist: A connectionist model of continuous speech recognition," *Cognition,* vol. 52, no. 3, pp. 189-234, 1994.

[5] D. Norris and J. M. McQueen, "Shortlist B: A Bayesian model of continuous speech recognition," *Psychological review,* vol. 115, no. 2, pp. 357-395, 2008.

[6] J. L. McClelland and J. L. Elman, "The TRACE model of speech perception," *Cognitive psychology*, vol. 18, no. 1, 1986.

[7] N. J. Smith and R. Levy, "The effect of word predictability on reading time is logarithmic," *Cognition*, vol. 128, no. 3, pp. 302-319, 2013.

[8] S. L. Frank, L. J. Otten, G. Galli, and G. Viliocco, "The ERP response to the amount of information conveyed by words in sentences," *Brain and language,* vol. 140, pp 1-11, 2015.

[9] R. M. Willems, S. L. Frank, A. D. Nijhof, P. Hagoort, and A. van den Bosch, "Prediction during natural language comprehension," *Cerebral Cortex,* vol. 26, no. 6, pp. 2506-2516, 2015.

[10] N. Oostdijk, "The design of the Spoken Dutch Corpus," *Language and Computers*, vol. 36, no. 1, pp. 105-112, 2001.

[11] R. van Son, W. Wesseling, E. Sanders, and H. van den Heuvel, "The IFADV Corpus: A Free Dialog Corpus," in LREC, pp. 501-508. Marrakech: ELRA, 2008.

[12] R. Schäfer, "Processing and querying large web corpora with the COW14 architecture," in *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3),* pp. 28-34. Mannheim: Institut für Sprache, 2015.

[13] R. Schäfer and F. Bildhauer, "Building Large Corpora from the Web Using a New Efficient Tool Chain," in *Proceedings of the Eight International Conference on Language Resources and Evaluation,* Istanbul: ELRA, pp. 486-493, 2012.

[14] W. Goedertier, S. M. Goddijn, and J. P. Martens, "Orthographic transcription of the Spoken Dutch Corpus," in *Proceedings of LREC-2000*. Athens: ELRA, 2000.

[15] A. Stolcke, "SRILM-an extensible language modelling toolkit," in *Proceedings of the International Conference on Spoken Language Processing*. International Speech Communication Association, 2002.

[16] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language*, vol. 13, no. 4, pp. 359-393, 1999.

[17] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely. "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.* The IEEE Signal Processing Society, 2011.

[18] R Core Team, "R: A language and environment for statistical computing," R Foundation for statistical Computing, Vienna, http://www.R-project.org/, 2015.

[19] D. Norris, J. M. McQueen, and A. Cutler, "Prediction, Bayesian inference and feedback in speech recognition," *Language, cognition and neuroscience*, vol. 31, no. 1, pp. 4-18, 2016.

[20] F. Huettig, "Four central questions about prediction in language processing," *Brain research,* vol. 1626, pp. 118-135, 2015.

[21] J. F. Connolly and N. A. Phillips, "Event-related potential components reflect phonological and semantic processing of the terminal word of spoken sentences," *Journal of Cognitive Neuroscience,* vol. 6, no. 3, pp. 256-266, 1994.

[22] A. Brunellière and S. Soto-Faraco, "The speakers' accent shapes the listeners' phonological predictions during speech perception," *Brain and language,* vol. 125, no. 1, pp. 82-93, 2013.