LIPS: a tool for predicting the lexical isolation point of a word

Andrew Thwaites[*], **Jeroen Geertzen**[†], **William D. Marslen-Wilson**[*] **and Paula Buttery**[†]

[*]The MRC Cognition and Brain Sciences Unit
15 Chaucer Road, Cambridge, CB2 7EF, UK
{andrew.thwaites, william.marslen-wilson}@mrc-cbu.cam.ac.uk

[†]Computation, Cognition and Language Group, RCEAL
English Faculty Building, 9 West Road, Cambridge, CB3 9DP, UK
{jg532, pjb48}@cam.ac.uk

Abstract

We present LIPS (Lexical Isolation Point Software), a tool for accurate lexical isolation point (IP) prediction in recordings of speech. The IP is the point in time in which a word is correctly recognised given the acoustic evidence available to the hearer. The ability to accurately determine lexical IPs is of importance to work in the field of cognitive processing, since it enables the evaluation of competing models of word recognition. IPs are also of importance in the field of neurolinguistics, where the analyses of high-temporal-resolution neuroimaging data require a precise time alignment of the observed brain activity with the linguistic input. LIPS provides an attractive alternative to costly multi-participant perception experiments by automatically computing IPs for arbitrary words. On a test set of words, the LIPS system predicts IPs with a mean difference from the actual IP of within 1ms. The difference from the predicted and actual IP approximate to a normal distribution with a standard deviation of around 80ms (depending on the model used).

1. Introduction

Recent theories in human speech perception assume that during the reception of acoustic evidence of a word, listeners activate and strengthen word candidates in the mental lexicon based on the word-onset being heard and recognise a word as soon as one of the candidates stands out sufficiently (Marslen-Wilson, 1987). The point at which a listener can reliably discriminate from other word candidates is known as the *isolation point*. For some words, especially monosyllables, this point may only be reached once the corresponding acoustic signal is complete. For other words, however, listeners are often able to recognise them well before the corresponding acoustic signal is complete (Marslen-Wilson, 1975). A listener may have reached the isolation point, marking the time from which the word is identified correctly, but may not be very confident about it. For this reason, some studies also identify the *recognition point* (RP). The RP is similar to the IP, but additionally requires the listener to be sufficiently confident about identification. [1] In this study, we will consider only IPs and ignore RPs.

The time at which lexical isolation occurs may be useful information in several kinds of studies, most notably in evaluating competing cognitive models of speech perception and in accurately aligning words with brain responses in neurobehavioural studies. Obtaining IPs for an arbitrary set of words, however, requires costly psycholinguistic experiments involving a fair number of subjects. In this paper we describe a predictive model that produces accurate isolation points for arbitrary speech-recorded words. This model is implemented in a tool that caters cognitive studies on speech perception.

2. The use of IPs

The accurate and automatic prediction of the isolation points of words is particularly useful in studies of speech production and the neuroscience of language. The acoustic and linguistic features that are used in IP prediction can be ranked according to their importance in the prediction task, providing valuable insights as to the kind of information that is important to the task. This in turn can provide valuable insights in studying human word and speech recognition. In neuroscientific studies of speech perception, accurate alignment of word-based stimuli with corresponding brain signals is important. Especially for brain-imaging techniques that offer high temporal resolution in the order of milliseconds, like magnetoencephalography (MEG) and electroencephalography (EEG), accurate word alignment improves analysis. Gating studies are often used to align words and signals, but such studies are rather time-consuming and require numerous behavioural tests for each stimulus.

3. Data

To train and test a predictive model for IP recognition, data are needed containing audio recordings of words and their corresponding isolation points. The IPs are usually obtained by *gating studies* (Grosjean, 1980). In a gating study, listeners are presented with increasingly long onsets of a word (e.g. /c/, /ca/, /cap/) and are asked what they think the word is, or going to become. The results indicate the minimal acoustic input needed to identify words in speech. For the work in this paper, gating data (audio recordings and timings) have been used that were gathered and studied by (Tyler et al., 2002). The dataset includes 160 words, each of which was presented to at least 10 subjects. The

[1] A reason why a listener may not be entirely confident is the activation of a few other candidates consistent with perception. A common criterion for ‘sufficiently confident’ is usually 80% confidence.

spoken words were recorded as complete words, and for each one 100 ms onset was presented, followed by onsets with increasing duration by 50 ms increments until the complete word had been heard. After each onset presentation, subjects were given five seconds to write down the word they heard. The IP was defined as the average gate duration at which a word was first correctly identified. A word was deemed to have been correctly identified when a participant had written the correct word and had not deviated from that response on subsequent trials. From the initial 160 words available, 6 words were discarded (these words were not recognised by the end of the audio file by all participants), leaving a dataset of 154 words, which all have a non word-final IP.

4. Word features

For the various models that are described in this work, several acoustic and linguistic word characteristics are used that may aid the prediction, listed in Table 1.

Table 1: Word characteristics used in prediction

features
absolute time through word (ms)
word length (ms)
phoneme probability
cumulative vowel tally
cumulative stressed vowel tally
cohort size (type)

The suggested characteristics in table 1 are all time varying (with the exception of word length). They come in two categories: 1) based on word frequencies (type and token) and 2) based on acoustic properties. In order to obtain acoustic properties (both phonemic and sub-phonemic) at high temporal resolution a speech recognition system is required. For this work, an automatic recogniser based on CMU Sphinx (Walker et al., 2004), has been built into the analysis module of the software. [2] The recogniser converts the acoustic signal into a probabilistic phoneme lattice and presents the most probable phoneme sequence, together with phoneme boundaries. Phoneme probabilities were obtained by converting the words in the frequency list of the British National Corpus (BNC) (Leech et al., 2001) to phoneme sequences using the Carnegie Mellon University Pronouncing Dictionary (Weide, 2008). In those features that require the recognition process, no *forced alignment* was used (a condition where the accuracy of the lattice can be improved when the output of the recogniser is known beforehand), so that the conditions under training and testing are the same. However, forced alignment for the training data may be found to increase the accuracy in future. The cohort size at time q is the number of words an audio file clipped between 0 ms and q ms could be starting. So the audio */ca/* could be beginning of the words */cap/*, */captain/* and */cat/*, among others. The number of possible completions (and thus the cohort size) at q ms will always drop as q increases, and will become one as soon as the phonetic dictionary or word list contains a single continuation. There are a number of ways to calculate this feature, but it was done here by finding the sequence of phonemes that have been heard up to time q using the speech recogniser, and then searching through a phonetic dictionary for all words that start with that phoneme sequence.

These features may have information that can be used to accurately predict the IP of a word. The following models try to use this information to generate accurate predictions of IPs for an unknown test set.

5. Predictive models

5.1. Model I

The first predictive model is a simple baseline model in which the relative position of the IPs in the training data are averaged. K is the model coefficient with a training set with n word-IP pairs:

$$K = \frac{1}{n}\sum_{i=0}^{n}\frac{IP_i}{duration(w_i)} \qquad (1)$$

where IP_i is the time in milliseconds between the isolation point and the onset of the audio file for a word-IP pair in the training set. The model coefficient is subsequently used in a linear function to predict the IP for word w:

$$IP(w) = K \cdot duration(w) \qquad (2)$$

This model uses only *duration* from the feature set.

5.2. Model II

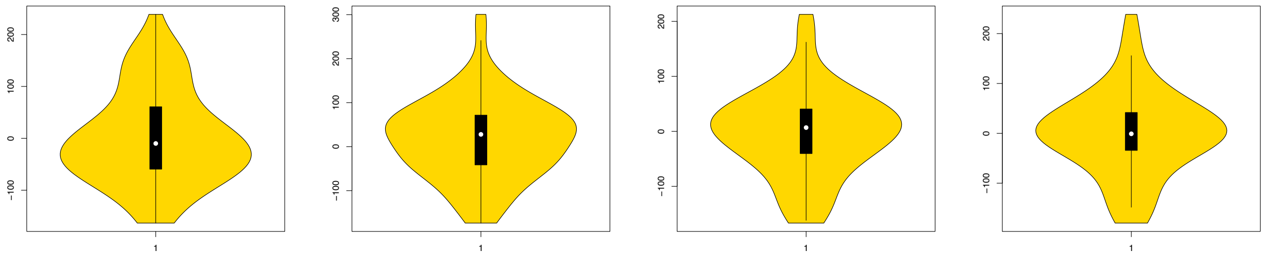
This model is based on the direct predictions that the cohort model (Marslen-Wilson and Tyler, 1980; Marslen-Wilson, 1987; O'Rourke and Holcomb, 2002) makes about the timing of the identification point. For isolated words, the cohort model assumes that full word identification occurs as soon as the word-onset is no longer compatible with multiple lexical candidates. The point at which the acoustic evidence allows the listener to single out a candidate is known as the *uniqueness point* (UP), which is expected to coincide with the observed IP. [3] To evaluate this model an audio recording was segmented at phoneme boundaries and the cohort size was calculated at each progressive phoneme. This model only uses cohort size as a feature to predict the isolation point, where cohort size was defined by process set out in section 4.

5.3. Model III

Predictive Model III uses logistic regression and takes an unspecified amount of word features from section 4 to predict the isolation point. To train this model the recordings were segmented at phoneme boundaries and the resulting

[2] Any other speech recognition toolkit, such as HTK (Young et al., 2006), could have been used for the analysis module.

[3] The UP and the IP need not correspond: the word may be recognised before a single candidate remains, if the context is helpful. It may also happen that there is a delay in isolating the word because of frequency effects.



(a) Model I: Baseline

(b) Model II: Cohort-model based

(c) Model III: Simple logistic regression

(d) Model IV: Ordinal logistic regression

Figure 1: Distributions of difference between the actual and predicted IPs

data set was split into two categories: before IP; and after IP. The logistic function is defined by:

$$f(Z) = \frac{1}{1+e^{-z}} \tag{3}$$

The variable z is representative of a set of predictors and is defined by:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \tag{4}$$

where β_0, β_1, β_2 ... β_k are the regression coefficients of predictors x_1, x_2 .. x_k respectively. The predictors used here were: absolute time from the onset of the word, word length, the type probability of the current phoneme, the number of vowels that have been encountered to that point, the number of stressed vowels that have been encountered up to that point and the size of the cohort at that point (calculated as in Model II). Model III was evaluated by calculating the value of the logistic function for predictors retrieved at successive phonemes in an unseen word and using the coefficients defined by the training data. The output of the logistic function is confined to values between 0 and 1. The time of the first phoneme achieving $f(z) > 0.5$ was taken to be the IP.

5.4. Model IV

Predictive Model IV uses ordinal logistic regression. The method and evaluation here is similar to that of Model III except that the training data is now split into 4 categories: word beginning; just before IP; just after IP; word ending.

6. Model evaluation

The models have been trained on the word-IP pairs using a leave-one-out strategy, with 10% of the data being retained for a test set. Model performance is assessed by inspecting the differences between the actual IP times (as specified by the gating experiment) and the predicted IP from the model. The mean, standard deviation and median of these differences can be seen in Table 2 and their distributions displayed pictorially in Figure 1.

Models I and II have a similar standard deviation in the values for the difference between predicted and actual IP time. However the 'differences distribution' for Model I is not normal and is in fact wider than that of Model II (see figures 1a and 1b). Model I (which bases predictions solely on relative distance through a word) is thus the worst performing model. Model II (which bases prediction on cohort size) performs slightly better since (despite being offset by a mean of 18ms) its predictive value is simpler to define. The performance of Models III and IV is reasonable similar. The spread of the differences is narrower than for Models I and II making them better IP predictors. However Model IV performs arguably best with a mean difference between actual and predicted IP of 0.44ms.

Table 2: Model performance metrics (sign signifies the direction of difference)

model	*mean*	*stdev*	*median*
I	2.16	90.14	-9.90
II	18.94	90.98	27.95
III	-2.31	77.91	6.80
IV	0.44	84.62	-1.03

7. Discussion

A limitation of the current models is that they are tuned for isolated words, and not for words in context. (Grosjean, 1980) noticed that a listener needed to hear an average of 199 ms of a word when it occurred in sentential context, as opposed to 333 ms for the same word presented in isolation. If we define the *uniqueness point* as the point at which there ceases to be any overlap with any other words in the mental lexicon, and we define the *isolation point* as the point at which the listener guesses the whole word, then a word may be recognised before there is one remaining candidate left because of its context. IPs for words in context will occur on average earlier than IPs for words in isolation, and to take advantage of word context and improve IP prediction accuracy, the next version of LIPS will incorporate a language model.

The training and performance of the speech recogniser is crucial for allowing accurate IP prediction. LIPS is currently trained on American English pronunciation (acoustic model and training dictionary), which will inherently lower the accuracy of IP prediction for other varieties of English. In future models, however, this will be amended to suit the stimuli under observation. Other improvements to LIPS will involve the use of various additional word and sub-word measures as features in the IP prediction, such as phoneme and cohort entropy scores.

8. Conclusions

This work confirms that IPs can be predicted using training-based models, which are shown to out-perform the naive baseline. This approach, once refined, will be of significant use to cognitive-researchers who rely on costly gating studies to inform their research.

Acknowledgments

This work is supported by EPSRC (grant EP/F030061/1) and the MRC (grants U.1055.04.002.00001.01 and G0500842).

Appendix I

Words used in the recognition experiments (table 3).

9. References

F. Grosjean. 1980. Spoken word recognition processes and the gating paradigm. *Perception and Psychophysics*, 28:267–283.

G. Leech, P. Rayson, and A. Wilson. 2001. *Word Frequencies in Written and Spoken English: based on the British National Corpus*. Longman.

William D. Marslen-Wilson and Lorraine K. Tyler. 1980. The temporal structure of spoken language understanding. *Cognition*, 8(1):1–71.

William D. Marslen-Wilson. 1975. Sentence perception as an interactive parallel process. *Science*, 189:226–228.

William D. Marslen-Wilson. 1987. Functional parallelism in spoken word-recognition. *Cognition*, 25(1-2):71–102, March.

Timothy B. O'Rourke and Phillip J. Holcomb. 2002. Electrophysiological evidence for the efficiency of spoken word processing. *Biological Psychology*, 60(2-3):121–150.

Lorraine K. Tyler, Helen E. Moss, Adam Galpin, and J. Kate Voice. 2002. Activating meaning in time: The role of imageability and form-class. *Language and Cognitive Processes*, 17(5):471–502.

R. L. Weide. 2008. *The Carnegie Mellon Pronouncing dictionary [cmudict 0.07a], School of Computer Science, Carnegie Mellon University,* `ftp://ftp.cs.cmu.edu/project/fgdata/dict/`.

S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland. 2006. *The HTK Book, version 3.4*. Cambridge University Engineering Department, Cambridge, UK.

Table 3: Words used in the recognition experiments

chief	drug	rake	trunk
hurricane	priest	garden	food
coat	bandit	storm	tart
swamp	alley	card	pencil
desk	ferry	window	pirate
blunder	motive	reason	law
temper	fate	thrill	benefit
cult	trend	custom	clamour
system	creed	total	pardon
fact	charm	loan	lust
soak	shine	cut	carve
thump	grab	stumble	fall
float	run	crush	fidget
sniff	slap	sprint	press
burst	hug	munch	offer
dare	stay	dodge	cope
amble	mix	abandon	recruit
need	assist	keep	rid
admire	say	mimic	hire
attempt	relax	leave	ramp
fan	bill	park	caravan
tusk	whisky	brush	crow
arrow	shop	cage	neck
temple	hut	kipper	coast
sofa	basket	falcon	affair
budget	trace	flow	value
gain	force	mercy	theme
bane	despair	issue	luck
harm	term	envy	fault
skill	clue	oath	skip
shriek	sing	chat	hop
jump	giggle	weep	pounce
shiver	tumble	clap	cuddle
hum	chew	scream	shake
hiss	roar/raw	rub	cling
demand	join	take	guess
lack	agree	get	astonish
insist	divide	accuse	make
let	publish	crave	ignore
believe	declare	attach	hope