1aSCb2. Unsupervised machine learning for the accurate classification of the discourse marker like in code-switching utterances

Page E. Piccinini* and Eric R. Kramer

*Corresponding author's address: Linguistics, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0108, ppiccinini@ucsd.edu

Spanish-English bilinguals use the discourse marker like in English, Spanish, and code-switching utterances. An acoustic analysis found that the [l] and diphthong in like is produced differently depending on the type of utterance in which it occurs. To investigate how exactly these differences are manifested in the acoustic signal, we built a logistic-polynomial regression model to classify like tokens based on acoustic data. The model first projects F1 and F2 values onto a space of time-dependent polynomials. We then apply multinomial logistic regression to classify these polynomials as English, Spanish, or code-switching. The area under the curve was 0.75, showing classification was significantly greater than random. This model outperforms a model that rely on static values for F1 and F2 and a model based on independent component analysis. The superiority of the polynomial model suggests that the time-dependent progression of F1 and F2 values, rather than absolute formant values, is useful for predicting an imminent code-switch. By building and comparing additional models to investigate other acoustic aspects of the signal we can be more informed when building perception experiments to see what aspects of the signal listeners are more likely to use anticipating an upcoming code-switch.

Published by the Acoustical Society of America through the American Institute of Physics

P. Piccinini and E. Kramer

© 2013 Acoustical Society of America [DOI: 10.1121/1.4799757]
Received 22 Jan 2013; published 2 Jun 2013
INTRODUCTION

In speech perception there are many aspects of the signal a listener must pay attention to in order to comprehend their conversational partner’s message. This can be all the more difficult for bilingual speakers, who often engage in code-switching, sometimes using both languages in a single utterance (e.g. Pfaff, 1979; Poplack, 1980; Myers-Scotton, 2008). This makes the task of processing incoming speech more complicated, as speakers have to anticipate a possible change in languages in addition to processing other information. This task would be easier if there were cues in the signal warning a listener of an upcoming switch, which listeners could use to prepare for an upcoming change in languages.

In spontaneous speech Spanish-English bilinguals use the discourse marker *like* in English, Spanish, and code-switching utterances. An acoustic analysis found that the [l] and diphthong in *like* is produced differently depending on the type of utterance in which it occurs (Piccinini, 2011). Spanish *likes* begin more backed and end higher and more fronted as compared to English *likes*. Code-switching *likes* both from English to Spanish and English to Spanish begin like the language preceding the switch, English or Spanish, and end somewhere in between both languages.

To investigate how exactly these differences are manifested in the acoustic signal, the present paper builds three models that use different methodologies to classify the tokens of *like*. Which model best classifies the tokens and teach us more about where differences lie between given categories. For example, is it the static differences at particular points in time that differentiate the tokens, or is it the general shape of the change over time that is important? Having a better understanding what defines one category and differentiates from another can be useful when designing perception experiments to see if listeners use these cues to anticipate code-switches.

METHODS

Tokens

All tokens of *like* were collected from a corpus of spontaneous speech between dyads of early Spanish-English bilinguals. A total of ten speakers’ productions (eight female, two male) were included. Each token was given one of four classifications: 1) English, 2) Spanish, 3) code-switch from English to Spanish (CS-ES), or 4) code-switch from Spanish to English (CS-SE). A token was only considered a code-switch if it occurred at the boundary between the languages. Example utterances are provided in (1) – (4).

1. English
   He would just act *like*, I don't know.

2. Spanish
   Me acuerdo uno es que *like* no sé quien.
   GLOSS: *I remember one that *like* I don't know who.*

3. CS-ES (immediately before a switch to Spanish)
   One of those barrels and *like* *estaba adentro*.
   GLOSS: *One of those barrels and *like* he was in it.*

4. CS-SE (immediately after a switch from Spanish)
   Yo me acuerdo que tenía que ir *like* before having mine...
   GLOSS: *I remember that I had to go *like* before having mine.*

Each token was segmented from the onset of the [l] to the offset of the diphthong. Formants (F1, F2, and F3) were measured starting 10ms from the start of the token (onset of the [l]) and ending 10ms from the end of the token (offset of the vowel) at 5 percent intervals. All formants were bark-transformed and normalized using the Bark Difference Metric (Syrdal & Gopal, 1986).

Statistical Models

Three logistic regression models were used to classify tokens: 1) standard logistic regression, 2) a polynomial-logistic hybrid regression, and 3) an independent component analysis (ICA) regression. All regressions included both F1 and F2 measurements.
All models were validated by 10-fold cross-validation. The four-way multinomial classification scheme (English, Spanish, CS-ES, CS-SE) was reduced to five binomial classifications (English vs Spanish, English vs CS-ES, Spanish vs CS-ES, English vs. CS-SE, Spanish vs CS-SE).

**Simple Linear Regression**

The classifications were regressed on bark-transformed formant measurements using the *glm* package in R.

**Polynomial Regression**

A third-order polynomial was fit to bark-transformed formant values for token using the *lm* function in R. Classifications were then regressed on the polynomial coefficients using the *glm* package in R.

**ICA Regression**

An independent component analysis was conducted on bark-transformed formant values. Classifications were regressed on the estimated source components.

**Model Comparison**

The performance of each model was evaluated using receiver-operator curves (ROC). ROCs plot the true positive rate (TPR) versus the false positive rate (FPR) for each model. TPR and FPR are defined below

\[
TPR = \frac{TP}{TP + FN} \\
FPR = \frac{FP}{FP + TN}
\]

where TP is the number of true positives, FP is the number of false positives, TN is the number of true negatives and FN is the number of false negatives. By varying the parameters of each model, we can generate a series of TPRs and FPRs for each model to build the ROC.

Roughly, better classifiers have curves closer to the top-left corner of the graph. A completely random classifier has a straight diagonal line for its ROC. The area under the curve (AUC) represents the area underneath the ROC and is frequently used to quantify the model as a whole. AUC values range from 0.5 (random) to 1 (perfect).

All ROC curves were built with the R package ROCR (Sing et al., 2005).

**RESULTS**

**Spanish Tokens**

The polynomial regression model was best at discriminating between English and Spanish tokens (area under the curve (AUC) = 0.72), followed by the ICA regression model (AUC = 0.72), and then the standard logistic regression model (AUC = 0.55). The difference in AUC was statistically significant for both the polynomial regression and ICA regression as compared to the standard logistic regression \( t(17.38) = -2.98, p < 0.01; t(15.80) = -3.65, p < 0.01 \). However, there was no statistically significant difference between the polynomial regression and ICA regression.
Figure 1. A receiver-operator curve for the categorization of Spanish tokens for the four models. The most accurate model (the highest AUC) is the polynomial model.

**Code-switching English-Spanish (CS-ES) Tokens**

For CS-ES tokens as compared to English tokens, the model with the highest AUC was the polynomial regression (AUC = 0.60), followed by the ICA regression (AUC = 0.58), and lastly the standard regression (AUC = 0.47). Unpaired t-tests found a significant difference between the standard regression and the polynomial regression \([t(14.50) = -3.12, p < 0.01]\) and a marginally significant difference with the ICA regression \([t(12.38) = -2.17, p = 0.05]\). There was no significant difference between the polynomial regression and the ICA regression.

Compared to Spanish tokens, the model with the highest AUC was the ICA regression (AUC = 0.64), followed by the polynomial regression (AUC = 0.54), and then the standard regression (AUC = 0.49). Unpaired t-tests found a marginally significant difference between the standard regression and the ICA regression \([t(17.97) = -1.83, p = 0.08]\). No other comparisons were significant.
Figure 2. Receiver operator curves for the categorization of CS-ES tokens for the four models as compared to English (a) and Spanish (b). The best model, as defined by AUC, is the polynomial model for the English comparison and the ICA model for the Spanish comparison.

Code-switching Spanish-English Tokens

Comparing CS-SE tokens to English tokens, the model with the highest AUC was the standard regression (AUC = 0.59), followed by the ICA regression (AUC = 0.53), and finally by the polynomial regression (AUC of 0.48). Unpaired t-tests found no significant differences between any of the models.

Comparing tokens to Spanish tokens, the model with the highest AUC was the ICA regression (AUC = 0.59), followed by the polynomial regression (AUC = 0.50), and followed by the standard regression (AUC = 0.46). Unpaired t-tests found no significant differences between any of the models.

Figure 3. Receiver operator curves for the categorization of CS-SE tokens for the four models as compared to English (a) and Spanish (b). No model was found significantly superior for either comparison.
DISCUSSION

The analysis in the present paper found that the polynomial regression was best at accurately classifying tokens of the discourse marker *like*. As such, it outperforms standard logistic regression, which is based solely on normalized formants, and an ICA regression model, which is based on a linear combination of estimated source variables. The superiority of the polynomial logistic held for almost all of the comparisons we tested. This suggests that it is the overall shape of the productions within a given category that is most informative for classification and not formant values at static time points or chunks of time times.

All three models were best at distinguishing Spanish from English tokens. Interestingly, the models had difficulty distinguishing code-switching tokens from both English and Spanish tokens. This may reflect the gradient nature of these productions, as previous research showed that code switching productions are in between the two languages. It is also possible that while discrimination between English and Spanish tokens is heavily influenced by F1 and F2, perhaps the distinction between code-switching and non-code-switching tokens relies upon other acoustic features, such as pitch or amplitude. Additional models using these components of the speech signal could be more informative as a way to classify tokens, and thus a clue into how listeners anticipate code-switches.

CONCLUSION

English and Spanish tokens of the discourse marker *like* can be accurately classified using polynomial regressions of F1 and F2 from the onset of the [l] to the offset of the vowel. Formants are not as informative in differentiating code-switching tokens from English or Spanish tokens, however. Future research should build additional models looking at other acoustic factors that may provide cues to the differences between code-switch and non-code-switch tokens. By building and comparing additional models to investigate other acoustic aspects of the signal we can be more informed when building perception experiments to see what aspects of the signal listeners are more likely to use anticipating an upcoming code-switch.

ACKNOWLEDGMENTS

Thank you to Amalia Arvaniti for reading over an earlier draft of this paper. Any and all errors are our own. We thank all the participants in this study.

REFERENCES