July 2016

The Language Identification Problem: Formant Analysis and Cross-Linguistic Uniqueness

Lyndon Rey
The University of Western Ontario, lrey@uwo.ca

Follow this and additional works at: https://ir.lib.uwo.ca/wpl_clw

Part of the Computational Linguistics Commons, Phonetics and Phonology Commons, and the Typological Linguistics and Linguistic Diversity Commons

Recommended Citation
Available at: https://ir.lib.uwo.ca/wpl_clw/vol3/iss1/5

This Article is brought to you for free and open access by the Linguistics Program at Scholarship@Western. It has been accepted for inclusion in Western Papers in Linguistics / Cahiers linguistiques de Western by an authorized editor of Scholarship@Western. For more information, please contact tadam@uwo.ca, wlswadmin@uwo.ca.
The Language Identification Problem: Formant Analysis and Cross-Linguistic Uniqueness

Cover Page Footnote
This work was completed under the supervision of Dr. Robert E. Mercer (University of Western Ontario, department of Computer Science). The data analyzed is drawn from the Heritage Language Variation and Change (HLVC) corpus, of which permission was graciously granted by Dr. Naomi Nagy (University of Toronto, department of Linguistics).

This article is available in Western Papers in Linguistics / Cahiers linguistiques de Western: https://ir.lib.uwo.ca/wpl_clw/vol3/iss1/5
The Language Identification Problem:
A phonetic-acoustic solution for identifying spoken language

Imagine, for a moment, the following scene: a student sits in a coffee shop, putting the finishing touches on a paper written for a computational linguistics class. Taking a break from his work, our student overhears two other students speaking in an unintelligible language. Although he has no problem with other students conversing in foreign languages, he is troubled by the fact that they continue to look directly at him, giggle, then continue the conversation; this cycle continues for several minutes. Our (rather curious) student finds himself thinking about how convenient it would be if he were to have an application on his cellphone that could identify the language being spoken, though based only on small snippets of conversation, and then translate the spoken words into his native tongue. Frustrated, the student returns back to his work, wondering whether (and why) people were gossiping about him.

Although this situation may seem rather light-hearted and menial, it opens the door to a much larger possibility. As languages are incredibly phonetically diverse, is it possible to identify any given language solely using nuances found in its respective phonetic characteristics? In the following work, a small part of this question will be answered; namely, that the vowel distribution of a language does show enough variation to correctly disambiguate a small subset of the world's languages. However, to begin, a background on
the components of a vowel must be examined.

Phonologically, every vowel is a phoneme (although the inverse is not necessarily true), wherein a phoneme is simply the smallest unit of sound that carries meaning (Chomsky & Halle, 1968). The definition of a phoneme is based in the principle of minimal pairs: If there exists two words in a language with a single phonemic alternation, then the two alternating phones (a phone is simply a sound) can be considered to change meaning, and therefore constitute phonemes (Swadesh, 1934). To use English as an example, consider the words 'tip' /tɪp/, and 'rip' /rɪp/. As the phones /t/ and /r/ are the alternating pair in otherwise phonetically identical (though semantically divergent) words, it can be assumed that they are both phonemes in English. As seen in Chomsky and Halle's archetypal work on phonology, The Sound Patterns of English (1968), (and later revised by Hayes (2009)) vowels can be characterized through a series of binary ‘features’, which abstractly represent the characteristics of any given phoneme. To exemplify this, consider the binary feature representation of the vowel [i] (as in the vowel in 'beep'):

\[
[i] = \begin{bmatrix}
+ & \text{Syllabic} \\
+ & \text{Front} \\
- & \text{Back} \\
+ & \text{High} \\
- & \text{Low} \\
- & \text{Round} \\
- & \text{Nasal}
\end{bmatrix}
\]

It must be noted that while the above characterization of [i] is relatively general, the system only requires enough features in order to uniquely identify the vowel in the given vowel inventory of the language in question. The main motivation behind a binary representation of phonemes using features is to allow generality, which in turn allows characterizations to made about any given language, and also allows cross-linguistic comparison.
Although this binary feature representation is effective in representing simple vowel systems, it falls short when considering that each vowel phoneme has a rather wide range of 'allowable' pronunciations (in this setting, 'allowable' refers to the functionality of the pronunciation – if the meaning of the word does not change when the pronunciation is varied, then it is, by definition, 'allowable'). Therefore, a deeper, acoustic structure must be considered. Any given vowel possesses a series of acoustic properties called 'formants'; these formants are simply the frequencies (in Hertz) at which the vowel harmonically resonates (Standards Secretariat, 1994). These formants can be seen in analyzing a spectrogram, where the formants are simple the darker bars of each vowel. It is also worth noting that the first formant $F_1$ corresponds to the height of a vowel, while the second formant $F_2$ corresponds to the backness of a vowel (the formants $F_3, F_4, \ldots, F_n$ modify timbre and can mark phonetic phenomena such as rhoticity, nasalization, and rounding) (Ladefoged, 2006).

Below (figure 1) is a spectrogram analysis of a small speech sample, where the red dots indicate formants (note how they correspond with the darker bars):

![Figure 1 - A Praat spectrogram showing formants (in red)](image)

For the purposes of this work, a vowel (as discussed in the previous two paragraphs) can be defined as follows: a unit of speech, that possesses $n$ formants and is the nucleus of a syllable (this is discussed in page 7). When referring to vowels in this work, the following notation will be used, which represents any given vowel as the column vector of $n$
dimensions, with $n$ being the number of formants:

$$V = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_n \end{bmatrix}$$

It is with this representation of a vowel as a point in $n$-dimensional space that the main issue of language identification can begin to be solved. Firstly however, the current methods of language identification must be examined.

Language identification can be split into two main subfields: spoken language recognition, and textual language recognition. For the purposes of this paper, the former will be examined most thoroughly; however, there are links between the two subfields. Textual language recognition is achieved mostly through the analysis of n-grams, such as in the work of Rehrurek & Kolkus (2009) – they use n-gram models matched with a probabilistic model based on training texts given to the program. Most other textual language identification algorithms use similar strategies, such as described by Dunning (1994), and Cavnar & Trenkle (1994). However, apart from the few spoken language identification algorithms which use language recognition to create an n-gram for input into the aforementioned probabilistic models, spoken language identification models tend to use phonetic data to disambiguate language.

The main spoken language identification model being considered is the work of Montavon (2009), who compares identification strategies of three languages (English, French, and German); this is an effective comparison with the novel research presented in this paper, as it also involves an acoustic-phonetic comparison of three languages. The source of Montavon's three corpora are from radio recordings, with about five radio stations for each of the three languages in question. Using a deep architecture structure for the analysis of his
training corpora, he achieves a language identification accuracy of 83.5% for new radio stations (ones not used as part of the training corpora), given a five second sample. In comparison with the following work, it must be noted that Montavon uses all data in each spectrogram; the work presented below only considers vowels.

To return to the theme of only considering vowels, a brief outline of the concept of vowel inventory must be considered. In any given language, there are certain vowels that are given phonemic status; they are considered to be phonemes, considering the principle of minimal pairs. There does not exist a language which has a vowel inventory using every vowel phoneme given by the International Phonetic Alphabet (there exists 29 vowels in the IPA phonemic representation of possible vowels (figure 2); when nasalization is considered (such as in French), many more vowel phonemes are possible). To exemplify a typical vowel inventory, figure 3 gives the vowel inventory of American English.

![Figure 2 - The IPA vowel chart](image)

![Figure 3 - Vowel Inventory for American English](image)

Although these phonemic representations of vowel inventories are general (and therefore not specific enough to compare cross-linguistically), they give a suitable frame for the expected
results of the following experiment.

Returning to the definition of a vowel (given on pages 4 and 5), vowels can be represented as a column vector of each measured formant, to the $n$-th formant. As any given language has a specific vowel inventory, 'peaks' are expected to be seen in a histogram analysis of any given sample – especially in the $F_1$ and $F_2$ formants, as they seem to be the most variable based on vowel inventory. For example, consider the histogram of $F_1$ (figure 4), created from the analysis of a ~25 minute sample of Italian speech. The corpus used for the following analysis – as well as the rest of the data in this paper – is the Heritage Language Variation and Change (HLVC) corpus, of which permission was graciously granted by Dr. N. Nagy (University of Toronto). The histogram in figure 4 was sampled using a sampling rate of 1/100 second, and only the specific samples containing vowels were considered; there are vowel 23773 samples in figure 4.

![Figure 4 - Italian F1 distribution](image)

The results seen in figure 4 are extremely promising, as the peaks seen in the distribution of frequencies (around 290Hz, 400Hz, 520Hz, 660Hz, and 730Hz) match the $F_1$ values of a
typical Italian vowel inventory (figure 5). The five peaks seen represent the most common vowels ([i e a o u]) respectively, with the other two ([ε ɔ]) requiring a combination of $F_1$ and $F_2$ before they are able to be distinguished ([ε] is rather close to [e], and [ɔ] is rather close to [o]).

Moving on to an examination of the data collection methodology, a large debt is owed to Dr. Paul Boersma and Dr. David Weenink, both from the Phonetic Sciences department at the University of Amsterdam. Their acoustic analysis software, Praat, was the computational tool used to extract phonetic data from the samples; for any given unit of time, an arbitrary number of formants (in Hz) are able to be sampled, as well as intensity (in dB), and pitch (in Hz). The method used to collect formant data was relatively straightforward – samples were taken at 1/100 second each, and then each sample's respective intensity (in dB) was examined. Using the concept that (in most circumstances) a vowel will be the nucleus of its respective syllable, and therefore have a peak in the intensity of the sample (Abe, Wilson, & Erickson, 2012). In figure 6, the intensity of the sample is marked by the yellow line. For the purposes of this work, only the samples taken above the baseline average of intensity were considered; this includes the peaks in the intensity measure, as well as samples near to it. As the typical input sample was ~25 minutes long, ~150,000 raw samples
were collected, and ~25,000 were above the required intensity threshold, and therefore were considered vowels (for the purposes of this work).

Once the samples were collected, they were tabulated based on formants. For this preliminary study, only the first three formants were measured and used. However, Praat has the ability to measure an arbitrarily large number of formants, though after about seven, it tends to become rather inaccurate. In further research, higher formants will most likely be sampled. Tabulation was done using Microsoft Excel, although in future work MATLAB will most likely be used instead, as it allows for more powerful and straightforward computation of frequency and in the creation of histograms. Once tabulation was complete, a histogram was created, such as seen in the following representations. Three histograms were created for each language, representing $F_1$, $F_2$, and $F_3$; as three languages (Italian, Korean, and Ukrainian) were examined, nine histograms were created in total. The frequency (x axis) scales decided in the histograms were based on the minimum and maximum formant values observed, as well as linguistically-possible values for vowels (for example, the $F_1$ value of a vowel will never be greater than about 1100Hz, etc. (Catford, 1988)). They are as follows:

**Italian**
The scales of the y axes in the above graphs vary, due to the range of frequencies given in the x axis. As higher formants are considered, the range of frequencies that are linguistically-possible for each formant grows considerably; the range of \( F_1 \) is \(-100\text{Hz}-1100\text{Hz}\), while the
range of $F_2$ is ~240Hz-2500Hz, et cetera.

Clearly, though examination of the above samples, each language shows some variation between the distributions of frequencies in each of the first three formants. This is a good result (as stated on pages 6 and 7), as each language in question has a specific vowel inventory, so the fact that the graphs are not incredibly similar proves that the methodology of the experiment is not flawed (at least not from the data collection aspect). This will be made clearer in the following pages, where the graphs will be compared.

Considering the main question of phonetic language identification, the following section will outline the theory being used to differentiate cross-linguistically. As seen in the histograms on pages 8 and 9, the distribution of frequencies in each formant seems to take a specific pattern based on the vowel inventory of the language in question. With this in mind, and if the histogram were to be translated into a probabilistic model of frequencies for the formant of the language in question (using some form of interpolated smoothing to create a continuous function, discussed on page 15), the probability that any given formant of a vowel sample from any language could be calculated using simple calculus and probability. A caveat: the previous statement makes the assumption that the formant distributions in the language being identified had already been 'learned' by the algorithm. The definition of a probability density function states that the area under a probability density function is equal to the probability that a random variable will be a constituent of the data set (Stirzaker, 2003). To put this in practical terms, the probability that a random variable will exist within a continuous range of $(a, b)$ is simply the definite integral from $a$ to $b$ of the probability density function $p(x)$. This is formalized by the equation, where $X$ is a random variable, and $S$ is a set of data:

$$P(X \in S) = \int_{a}^{b} p(x) \, dx$$
To put this in terms of language identification, the probability that a measured formant $f$ will be a part of any given language $L$, assuming a probability density function respective to the measured formant $p_f(f)$, can be demonstrated by the following equation:

$$P(f \in L) = \int_a^b p_f(f) \, df$$

where and $a$ and $b$ represent the interval of precision measured. The interval of precision simply refers to the range of frequencies for which the probability is being calculated. In other words, the interval of precision simply refers to the probability that the given sample is within $a$ and $b$, in Hz.

Furthermore, it must be considered that within each sample, there exists an arbitrarily large number of formants, based on the number measured for the purposes of the study. As the probability of related events is cumulative (Stirzaker, 2003), the probability that any given sample slice of an audio input of unknown language will be the product series of each of the probabilities that each formant exists within any given language. Plugging this into the above equation, where $S$ represents a sample of arbitrary length, within a larger audio input, the following equation is formed:

$$P(S = L) = \prod_{f \in S} \left( \int_a^b p_f(f) \, df \right)$$

Based on this equation, it can be concluded that the accuracy of the probability calculated will increase conversely with the number of formants being measured. This holds true until a certain number of formants are considered, at which the accuracy of the tool used to measure begins to fail (with Praat, this is around seven formants), causing less accurate probabilities to be multiplied into the cumulative probability, therefore lowering the total accuracy of the algorithm.
Finally, the probability that an audio input $I$ of $S$ samples will exist within any given language $L$ is simply the product series of the above equation, for each sample within the audio input. This equation can be notated as follows:

$$P(I = L) = \prod_{S \in I} \left( P(S \in L) \right)$$

thus, and as $P(S \in L)$ is given by the second equation on page 11:

$$P(I = L) = \prod_{S \in I} \left( \prod_{f \in S} \left( \int_{a}^{b} p_f(f) \, df \right) \right)$$

can be used to model the probability that any given audio input will be a certain spoken language.

Also, there is the chance that through practical analysis of the above equation, the probabilities will be too low to draw meaningful conclusions, as outlying sample values (with very probabilities) could skew the probability density function. In this case, the probability for each sample could still be calculated using the equation at the top of this page. However, instead of creating a product series with these probabilities for each sample, a 'winning' language could be declared for each sample (the 'winning' language would simply be the most probable one), and then whichever language had the largest number of wins once all the samples were tested, could be declared the most likely language based on the audio input. For example, assume Italian had the highest probability for $i\%$ of samples, Korean had $k\%$ of highest samples, and Ukrainian had $u\%$ of highest samples. If $i > k$ and $i < u$, then Italian is the 'winning' language, and therefore would be the guess made by the algorithm considering the identity of the language. Another, (rather more simplistic) solution to this problem is to simply ignore the outlying data; however, if this strategy in taken, then there is a chance that subtle nuances in the formant distributions of a language could be missed.
Moving on to an analysis of data measured using the above method, a comparison of Italian versus Korean and Korean versus Ukrainian was performed. For the comparison, the histograms of the formants in each respective language were simply superimposed. For the purposes of this analysis, a positive result would be to observe a divergence between the formant distributions for each respective language, thus allowing probabilistic language identification to be performed. For instance, the comparison of $F_3$ (KOR vs. ITA) is highlighted, as it shows the largest divergence.

**Italian vs. Korean**

![Graphs showing comparison of Italian vs. Korean for F1, F2, and F3 formants.](image)

In the above results, $F_1$ and $F_2$ show a large amount of similarity – their respective plots are nearly identical. This result was unexpected preliminarily, as the (incorrect) assumption that
two genealogically-unrelated languages would show markedly different formant (and therefore vowel) distributions was made. However, in this specific case, Korean and Italian actually show rather similar vowel distributions, hence the similar $F_1$ and $F_2$ distributions. Notably, both languages do not possess horizontally-central vowels in their respective phonemic inventories; due to this, $F_2$ (which models horizontal articulatory position) shows two distinct 'bumps', representing the front (higher frequency bump) and back (lower frequency bump) vowels. Although this similarity does not directly support the main hypothesis, it proves that the methodology is accurate. Based on the known vowel distributions for Korean and Italian (Dryer & Haspelmath, 2013), two languages with similar vowel systems should show similar $F_1$ and $F_2$ distributions, regardless of genealogical relationships.

When $F_3$ is considered, the results become more interesting. In the above graph, there is clear bump in the Italian distribution at ~2500Hz, while the Korean distribution has a clear depression at the same frequency. Furthermore, the peak in the Italian distribution (at ~2950Hz) does not reach as high as the distribution seen in Korean, also at ~2950Hz. Although research has not of yet shown a conclusive reason behind this variation, it is a promising result; nuances in language-specific distribution of the same formant, as seen in $F_3$, are markers that could potentially used to disambiguate between language. These marked variations between formant distributions become even more obvious when two languages with dissimilar vowel inventories are considered, such as Korean versus Ukrainian:

*Korean vs. Ukrainian*
In the above comparative histograms, there is clear variation between each formant of each respective language. Ukrainian, with a very evenly-distributed vowel inventory (Dryer & Haspelmath, 2013), shows smooth curves (nearly a normal distribution, especially in $F_1$ and $F_3$), while Korean, as analyzed above, shows a clustered distribution.

Based on the above data, it is clear that any given language will show specific distributions in each formant. Through the interpolation strategy Kernel Density Smoothing (KDE) (performed in MATLAB), a function could be created, and applied to the previously-outlined probability distributions. From this, cumulative probabilities for a sample of arbitrary length could be calculated; the accuracy of these probabilities should positively correlate with an increased number of formants sampled, as well as the length of the audio input (as more viable vowel samples would be available). Once a probability of a certain language became higher than an arbitrary accuracy, that language could be declared the 'winner', as the most probable language of input. Considering the accuracy of the formant probability density functions, further refinement will occur once more data is processed. In other words, with a higher number of samples in the distribution histogram, the probability density function will model the learned language more realistically. Although the research presented in this paper is very preliminary, it shows good promise – language appears to show measurable nuance in formant distribution. Using these patterns, it is conceivable (given more data) that any spoken
language could be positively identified through phonetic variation alone.
References:


