INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

# Spanish Phone Confusion Analysis for EMG-Based Silent Speech Interfaces

*Inge Salomons*[1], *Eder del Blanco*[1], *Eva Navas*[1], *Inma Hernáez*[1]

[1]HiTZ Basque Center for Language Technology, University of the Basque Country, Spain

inge.salomons, eder.delblanco, eva.navas, inma.hernaez@ehu.eus

## Abstract

This paper describes a set of phone classification experiments based on electromyography (EMG) signals and a subsequent phone confusion analysis, as part of a project that aims to restore speech for Spanish laryngectomees by developing a Silent Speech Interface (SSI). Understanding the relationship between speech and the muscles used for speaking is essential to learn the possibilities and limitations of such EMG-based SSIs, before advancing to a complex task such as direct EMG-to-speech conversion. When considering only information from the muscles of the face and neck, important information from the tongue and vocal cords is missing. This is reflected in the results, which show confusion between pairs of phones that only differ in the position of the tongue or the voicing feature.

**Index Terms**: speech recognition, human-computer interaction, silent speech interfaces, EMG signals, speech processing, phone classification

## 1. Introduction and Related Work

When people lose their ability to produce speech naturally, alternative methods to communicate arise. One example is the case of laryngectomees, whose larynx, including vocal cords, were removed as part of cancer treatment and often re-learn to speak using an electrolarynx or by means of esophageal speech. However, these alternative methods are difficult to learn, and the resulting voice can be difficult to understand by others [1, 2]. For this reason, other technological alternatives, such as a Silent Speech Interface (SSI) [3], are being explored. SSIs take biosignals, such as electromyographic (EMG) signals, generated while speaking silently and output synthetic speech [4, 5]. Silent speech refers to the act of articulating without producing any sound, but in order to train an SSI model, often audible speech is used. The ideal EMG-based SSI converts EMG signals to speech in real-time in a fast and efficient way. Attempts at EMG-to-speech conversion have been made [6, 7, 8, 9, 10], but there is still room for improvement. We argue that it is important to gain information about the possibilities and limitations of such SSI´s by looking closely at all the individual steps that are involved. Examples of tasks that can provide useful information are syllable identification [11], word recognition [12], and speaker identification [13, 14]. The task that we focus on in this paper is phone classification [15, 16, 17], with the aim to analyze phone confusion after using EMG signals to classify phones and to gain insight into the relationship between muscle movement and speech. Previous phone confusion analysis for English showed that detecting voice as well as the manner of articulation is challenging [16]. Our study focuses on the Spanish language, but we expect a similar trend since surface EMG electrodes are located in the face and neck, which makes capturing the inner movement of the tongue difficult, and the tongue is an important part of speech production in either language. However, Spanish uses a different phone set, so we believe that for the development of an EMG-based SSI for Spanish a language-specific phone confusion analysis is necessary because it could provide new insights.

The outline of this paper is as follows. In Section 2.1, we describe the database used in this study and the data processing procedure. Sections 2.2, and 2.3 explain the feature extraction and reduction methods. Section 2.4 describes the model architecture and experiments. In Section 3, the results of the experiments are presented. These results are analyzed and discussed in Section 4. Finally, in Section 5, we provide a summary of the findings.

## 2. Method

The methodology consists of a classification task aiming at predicting the correct phone label for EMG frames. We used a database of synchronized speech and EMG data and trained a one-layer feed-forward neural network using cross-validation with features extracted from those signals.

### 2.1. Data

#### *2.1.1. The database*

The database consists of 28 sessions recorded by six native Spanish speakers (3 males and 3 females) aged 29 to 61, with a total of 11.5 hours of audible speech data. The number of recorded sessions differs per speaker (see Table 1 for an overview of the sessions and the total audio duration per speaker). In each session, a consistent base set of utterances was recorded, consisting of three distinct sets: 110 VCV combinations, 100 isolated words, and 100 sentences. Additionally, each session included another set of sentences, which was unique for each session, but remained consistent across speakers. During data processing, each set of words or sentences was split into an 80% train set and a 20% test set. This division process was applied uniformly to each session to ensure consistency. It is important to note that the test set was derived from the 100 sentences within the base set. As a result, the utterances in the test set remained the same across all sessions. Recordings are still ongoing and the database, which was approved by an ethics committee, will be publicly available when finished.

The EMG signals were recorded with a Quattrocentro amplifier (with a sampling frequency of 2048 Hz) using an eight-channel bipolar electrode setup targeting the following eight muscles: levator labii superioris, masseter, risorius, depressor labii inferioris, zygomaticus major, depressor anguli oris, anterior belly of the digastric, and stylohoid. The locations of the



10.21437/Interspeech.2023-1881

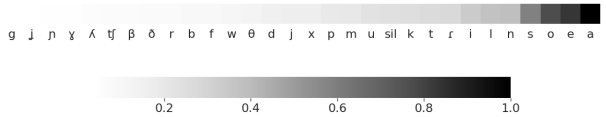


Figure 1: *Frequency of labels in the database, normalized with respect to the most frequent label, [a]. The darker the shade, the higher the relative number of frames labeled with that label.*

electrode pairs were determined after performing a pilot study, which focused on previous work done in this area of research, a study of the physiology of facial muscles, and experiments to compare the impact of several locations in the face and neck. The corresponding audio signal and a synchronization signal to align EMG and speech are recorded with the EMG signals.

Table 1: *Overview of the database: speakers, sessions and total audio duration (hh:mm:ss) per speaker.*

| Speaker | Sessions | Total Duration |
|---|---|---|
| 1 | 1-5, 7, 8 | 2:17:28 |
| 2 | 1-5, 7, 8 | 3:11:54 |
| 3 | 1, 2 | 1:00:02 |
| 4 | 1, 2, 5 | 1:09:26 |
| 5 | 1-3, 5 | 1:32:39 |
| 6 | 1-5 | 2:28:28 |
| all | | 11:39:57 |

#### 2.1.2. Data processing

The synchronization signal to ensure alignment between the EMG and audio signals is generated by the signal amplifier at the beginning and at the end of each utterance and is registered into an extra EMG channel. The synchronization signal is also outputted by the signal amplifier as an analog signal and introduced into one of the audio channels. This ensures that a mono speech recording is contained in one channel and the synchronization signal is stored in the other.

The first step in the data processing is to crop the EMG and the audio signals using the synchronization signal. Then, the audio is segmented using the Montreal Forced Aligner [18] to obtain the labels that are used to perform the classification. The phonetic transcriptions are obtained with a public transcriber[1], using the SAMPA phone set, which consists of 29 phones for Spanish. In this paper, we refer to the phones using the International Phonetic Alphabet (IPA) for more clarity. The silences at the start and the end of every utterance are discarded, but the short pauses between words are kept (represented as 'sil'). The distribution of the labels is shown in Figure 1.

### 2.2. Feature extraction

Instead of using the raw EMG and audio signals, we extracted features from them to use as input data.

#### 2.2.1. EMG-TD features

To parameterize the EMG signals, we calculated a set of Time Domain (TD) features, which have been shown to be effective in previous works [19, 16, 20]. In order to obtain these features, first we pre-processed the EMG signals by removing the continuous component and normalizing them. We calculated a nine-point double-averaged signal ($w$), a signal resulting from subtracting $w$ from the original signal ($p$), and a signal resulting from rectifying the $p$ signal ($r$). We then used a rectangular window of 25 ms duration and 5 ms step size to calculate the mean and power of $w$, the mean and power of $r$, and the zero-crossing rate of $p$, which are the 5 TD features that will characterize each frame.

The addition of temporal context information is essential as the signals related to the movement of muscles are not necessarily simultaneous to the generated speech. This means that relevant information might not be in the central frame but in the surrounding frames. To incorporate temporal context information into each frame, a stacking filter is applied, which allows to combine the features of the current frame with those of adjacent frames. We chose a stacking filter width of 15 frames, which means that the actual frame is stacked with the 15 preceding frames and the 15 subsequent frames, so that information from a total of 31 frames (i.e. 135 ms) is used. This results in a high-dimensional feature vector for each frame, with a length of $n \cdot 5 \cdot (2k+1)$, where $n = 8$ is the number of EMG channels and $k = 15$ is the width of the stacking filter, resulting in a total of 1240 features per feature vector.

#### 2.2.2. Audio MFCCs

To perform phone classification using acoustic features, we computed Mel Frequency Cepstral Coefficients (MFCCs) using a Hamming window and a filterbank of 30 filters, calculating 13 features for each window. The window length and the frame shift were identical to those used for TD feature extraction. As with the EMG features, we applied a stacking filter with a width of 15 to each frame, resulting in a feature vector with 403 audio features.

### 2.3. Feature reduction

To reduce the dimension of the feature vector, we applied Linear Discriminant Analysis (LDA) [21], which was demonstrated to be effective in previous studies [20, 16]. The maximum number of features allowed for LDA is the number of classes minus 1, which is in this case 29, since there are 30 phone classes. In order to find the optimal number of LDA features, we took session 2 of each speaker and performed a simple classification task on the EMG data following the model architecture described in Section 2.4.2, using a batch size of 64 samples and 10 epochs and iterating over 1 to 29 features. See Figure 2 for the validation accuracy of each feature averaged over all speakers. Based on this graph, we chose 17 features for all experiments described further in this paper.

### 2.4. Model architecture

#### 2.4.1. Baseline

To function as a baseline, a dummy classifier is used to achieve chance-level accuracy. This dummy classifier chooses the most common class. Due to the unbalanced label distribution as shown in Figure 1, using a baseline that represents random selection (in this case 3.33%) would not be fair.

#### 2.4.2. Feed-forward Neural Network

The neural network used to perform the phone classification consists of one feed-forward hidden layer with 34 nodes (double

[1] https://github.com/aholab/AhoTTS

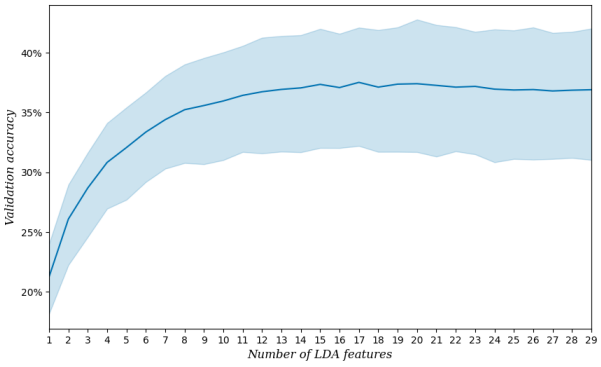


Figure 2: *Validation accuracy (in %) per number of LDA features averaged over session 2 of each of the six speakers. The solid line represents mean accuracy, and the area above and below the line shows the standard deviation range.*

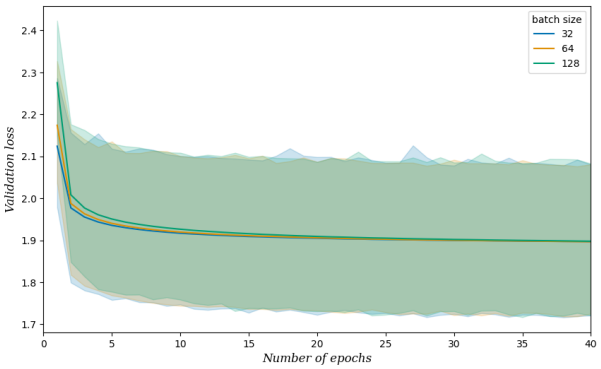


Figure 3: *Validation loss per number of epochs averaged over session 2 of each of the six speakers, using a Simple Feed Forward Neural Network after LDA reduction with 17 features. Results are shown for three different batch sizes. The solid line represents the mean validation loss, and the area above and below the line shows the standard deviation range.*

the number of inputs), and an output layer with 30 nodes (the number of phone classes). The activation function for the hidden layer is ReLU, while the output layer has a softmax activation function. As a metric to measure the multi-class classification accuracy of the network, the categorical cross-entropy loss function is used. The network was trained using an Adam optimizer and a learning rate of $10^{-3}$. For a similar task in previous work [17], we have also compared other classification models, such as bagging and Gaussian Mixture Models (GMMs). We learned that when using smaller datasets, a bagging classifier outperforms a neural network, but that a neural network is more effective when working with larger datasets.

We performed a hyper-parameter search in order to find the optimal batch size and number of epochs. Figure 3 shows the validation loss for 40 epochs and three different batch sizes: 32, 64 and 128. It can be seen that the batch size has no significant effect, so we chose 128, which has the shortest training time. Since the learning curve flattens after about 20 epochs, we chose this number for the final configuration.

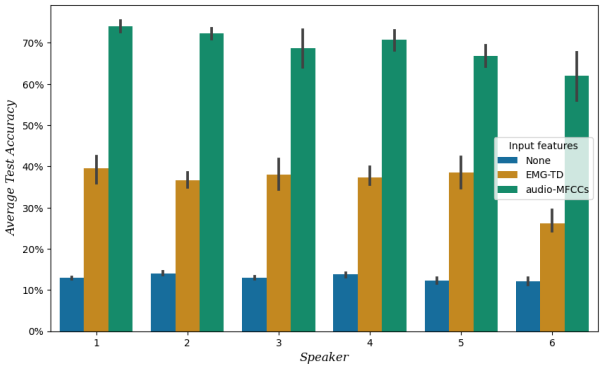


Figure 4: *Test accuracy averaged over sessions per speaker for different types of input features: none (most common class), EMG-TD and audio-MFCCs. The solid lines represent mean accuracy, and the areas between the dashed lines show the standard deviation ranges.*

### *2.4.3. Cross-validation*

As described in Section 2.1.1, we split the data into 80% train and 20% test sets. The test accuracy mentioned further in this paper represents results based on the test set. However, in the training phase, we used a cross-validation procedure using 5 folds, and the average accuracy of these folds is referred to as the validation accuracy.

## 3. Results

The mean validation and test accuracy for all experiments are shown in Table 2. All the accuracy values mentioned in this paper are frame-based. For the results per speaker, see Figure 4. The average time per model run was 13 minutes per session (without considering the dummy classification experiments). It can be seen that there is some variation between speakers, especially between speaker 6 and the other speakers. When speaker 6 is not taken into account, the mean test accuracy based on EMG features increases to 38.12%

Table 2: *Validation (including standard deviation) and test accuracy for all experiments, averaged over all speakers and sessions.*

| Input features | Validation accuracy | Test accuracy |
|---|---|---|
| None | 13.86±1.17% | 13.10% |
| EMG-TD | 37.52±5.34% | 35.98% |
| Audio-MFCCs | 67.03±6.75% | 69.68% |

Figure 5 shows the confusion matrix for all phone classes for the classification on the test set of EMG features. The matrix is normalized by the true labels, to account for the imbalance of phone classes. The matrix is organized by a shared phonetic feature, namely the manner of articulation, resulting in the following phone groups: vowels ([a], [e], [i], [o], [u]), semivowels ([j], [w]) and consonants. The consonants are further divided into plosives ([b], [p], [t], [d], [k], [g]), fricatives ([β], [f], [θ], [ð], [s], [x], [ɣ]), affricates ([tʃ], [ʝ]), nasals ([m], [n], [ɲ]) and liquids ([l], [ʎ], [ɾ], [r]). The label 'sil' refers to the short pauses in between words.

Table 3 shows the phone confusion pairs within each group,

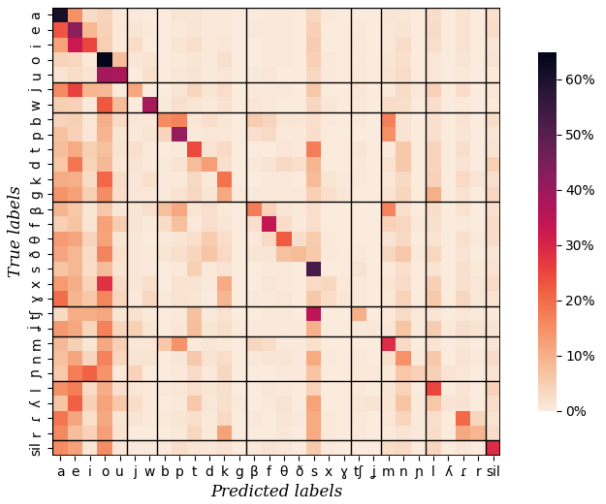


Figure 5: *Confusion matrix of the results of the test sets averaged over all speakers and sessions for the EMG features.*

of cases where the confusion was higher than the accuracy of the true label.

The silences were predicted correctly in 28.94% of the cases.

Table 3: *Table of within-phone group confusions, showing instances where the confusion between phones was higher than the accuracy of the true phone.*

| True | Accuracy | Predicted | Confusion | Group |
|---|---|---|---|---|
| i | 25.18% | e | 32.86% | vowels |
| u | 37.57% | o | 38.68% | |
| ɲ | 4.50% | n | 6.13% | nasals |
| ʎ | 1.41% | l | 6.32% | liquids |
| | | ɾ | 1.57% | |
| r | 9.02% | ɾ | 11.31% | |
| b | 15.67% | p | 16.49% | plosives |
| g | 0.57% | k | 11.39% | |
| | | t | 3.43% | |
| | | p | 1.72% | |
| | | d | 1.13% | |
| ɣ | 0.95% | s | 5.93% | fricatives |
| | | x | 2.35% | |
| | | θ | 1.00% | |

## 4. Discussion

In the previous section, we presented the results of our phone classification experiments. To start with, they showed that the mean chance level is 13.10%, and the mean accuracy based on EMG features is 35.98% (see Table 2), which implies that phones can to some extent be differentiated using information from the muscles. We also presented the results of the exact same experiment but this time using features from the audio signals, which lead to a mean accuracy of 69.68%. This result validates the model architecture but is not used for analysis since it does not contribute to the goal of this study.

Since the mean accuracy does not provide information about the phones individually, we show more detailed results about phone confusion. We found a Pearson´s correlation between the phone accuracy and label counts of 0.79. We can observe in Figure 5 that the phones [a], [e], [o] and [s] are predicted more often than other phones, to be recognized as vertical lines. As can be seen in Figure 1, these phone classes are the ones with the highest counts, so this pattern is most likely an effect of the correlation. Using the same reasoning, we can observe that those labels whose presence is rare, like [ɣ], [ʝ], [x] or [ʎ], are almost never predicted, what can be recognized as white columns.

When two phones that show confusion are members of the same phone group, this confusion can be explained in terms of their phonetic features. For example, as can be seen in Table 3, the vowel [i] is more often predicted incorrectly as [e] than correctly as [i]. Similarly, the vowel [u] is more often predicted incorrectly as [o] than correctly as [u]. These vowel pairs are indeed very close in their manner of articulation and the difference in muscle movements is subtle enough to explain this confusion. From the nasals group it is not surprising that the [n] and the Spanish [ɲ] show some level of confusion since the biggest difference between those two phones lies with the movement of the tongue, which is hard to capture with EMG. The same holds for the two different r´s in Spanish, the [ɾ] and the more tongue-rolling [r], and the two different l´s, within the group of liquids. When looking at the plosives, two unvoiced-voiced pairs ([p]-[b] and [k]-[g]) show confusion among them, which is as expected, since they only differ in the voicing feature, and the EMG electrodes are unlikely to pick up on that. They also share the place of articulation, so there is very little phonetic difference between them. Similar confusion between voiced-unvoiced phone pairs was also reported in previous work [16].

## 5. Conclusion

The analysis of the classification accuracy of phones based on EMG signals shows that it is possible to derive certain information from them, yet the results also revealed some level of phone confusion. More specifically, confusion between two phones is more likely to occur when they share the manner of articulation, or only differ in voicing. We are confident that part of this issue can be addressed in higher-level applications such as EMG-to-speech and EMG-to-text by applying language models, which we are currently working on. In general, we have found that acquiring good-quality EMG data is a challenging task. During the recording procedure, we encountered some technical issues such as electrodes detaching from the skin due to sweat and long recording sessions, or incorrect synchronization signals. Future research also includes enhancing the recording procedure, in order to reduce the technical issues and potentially improve the accuracy of the analysis.

## 6. Acknowledgements

This work has been funded by Agencia Estatal de Investigación ref.PID2019-108040RB-C21/AEI/10.13039/501100011033. We would like to thank all participants in this study.

## 7. References

[1] B. Weinberg, "Acoustical properties of esophageal and tracheoesophageal speech," *Laryngectomee rehabilitation*, pp. 113–127, 1986.

[2] T. Most, Y. Tobin, and R. C. Mimran, "Acoustic and perceptual

characteristics of esophageal and tracheoesophageal speech production," *Journal of communication disorders*, vol. 33, no. 2, pp. 165–181, 2000.

[3] J. Freitas, A. Teixeira, M. S. Dias, S. Silva *et al.*, *An Introduction to Silent Speech Interfaces*. Springer, 2017.

[4] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[5] J. A. Gonzalez-Lopez, A. Gomez-Alanis, J. M. Martín Doñas, J. L. Pérez-Córdoba, and A. M. Gomez, "Silent speech interfaces for speech restoration: A review," *IEEE Access*, vol. 8, pp. 177 995–178 021, 2020.

[6] M. Zahner, M. Janke, M. Wand, and T. Schultz, "Conversion from facial myoelectric signals to speech: a unit selection approach," in *Interspeech*, 2014, pp. 1184–1188.

[7] L. Diener, M. Janke, and T. Schultz, "Direct conversion from facial myoelectric signals to speech using deep neural networks," in *2015 International Joint Conference on Neural Networks (IJCNN)*, 2015, pp. 1–7.

[8] M. Janke and L. Diener, "EMG-to-Speech: Direct Generation of Speech From Facial Electromyographic Signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2375–2385, 2017.

[9] D. Gaddy and D. Klein, "Digital voicing of silent speech," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020, pp. 5521–5530.

[10] H. Li, H. Lin, Y. Wang, H. Wang, M. Zhang, H. Gao, Q. Ai, Z. Luo, and G. Li, "Sequence-to-sequence voice reconstruction for silent speech in a tonal language," *Brain Sciences*, vol. 12, no. 7, p. 818, 2022.

[11] E. Lopez-Larraz, O. M. Mozos, J. M. Antelis, and J. Minguez, "Syllable-based speech recognition using EMG," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, 2010, pp. 4699–4702.

[12] M. Wand and J. Schmidhuber, "Deep Neural Network Frontend for Continuous EMG-Based Speech Recognition," in *Interspeech*, 2016, pp. 3032–3036.

[13] L. Diener, S. Amiriparian, C. Botelho, K. Scheck, D. Küster, I. Trancoso, B. W. Schuller, and T. Schultz, "Towards Silent Paralinguistics: Deriving Speaking Mode and Speaker ID from Electromyographic Signals," in *Interspeech*, 2020, pp. 3117–3121.

[14] M. U. Khan, Z. A. Choudry, S. Aziz, S. Z. H. Naqvi, A. Aymin, and M. A. Imtiaz, "Biometric Authentication based on EMG Signals of Speech," in *2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, 2020, pp. 1–5.

[15] Q. Zhou, N. Jiang, and B. Hudgins, "Improved phoneme-based myoelectric speech recognition," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 8, pp. 2016–2023, 2009.

[16] M. Wand and T. Schultz, "Analysis of phone confusion in EMG-based speech recognition," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 757–760.

[17] E. D. Blanco, I. Salomons, E. Navas, and I. Hernáez, "Phone classification using electromyographic signals," in *IberSPEECH 2022*. ISCA, Nov. 2022, pp. 31–35.

[18] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[19] S.-C. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel, "Towards continuous speech recognition using surface electromyography," in *Ninth International Conference on Spoken Language Processing*, 2006.

[20] M. Wand, *Advancing electromyographic continuous speech recognition: Signal preprocessing and modeling*. KIT Scientific Publishing, 2015.

[21] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.