# Perceptual Linear Prediction Feature Extraction For Lampung Voice to Text Transcription (Case Study: Lampung Pepadun Dialect A)

Mohamad Zainudin[1], Akmal Junaidi[2*], Favorisen R Lumbanraja[3], Aristoteles[4], Tristiyanto[5]

[1,2,3,4,5] Faculty of Mathematics and Natural Science, Department of Computer Science, University of Lampung, Bandar Lampung, Indonesia
mohamadzainudin05@gmail.com
akmal.junaidi@fmipa.unila.ac.id
favorisen.lumbanraja@fmipa.unila.ac.id
tristiyanto.1981@fmipa.unila.ac.id

**Abstract.** Lampung language is one of the endangered regional languages in Indonesia, with the number of native speakers continuing to decline. One of the efforts to preserve this language is to utilize Speech-to-Text (STT) technology, which enables voice-to-text transcription to efficiently document the language. This research uses the Perceptual Linear Prediction (PLP) method for feature extraction from the voice signals of native speakers of Lampung Pepadun Dialect A. Perceptual Linear Prediction (PLP) was chosen for its ability to mimic human auditory perception, so that it can produce acoustic features suitable for the speech recognition process. The extraction process was performed on voice data from four speakers who uttered sentences in Lampung. The extraction results show that the PLP method is effective in capturing variations in acoustic characteristics, such as intensity, mid-frequency, and pronunciation stability. Each Perceptual Linear Prediction (PLP) coefficient provides significant information about the relevant sound elements for recognition. With these results, the developed Speech-to-Text (STT) system has the potential to be used in Lampung language preservation efforts through digital documentation and automatic transcription.

**Keywords:** Perceptual Linear Prediction (PLP), Speech-to-Text (STT), Lampung Language, Language Preservation, Feature Extraction.

## 1 Introduction

Lampung is one of the regional languages in Indonesia encompassing two main dialects, Lampung Pepadun Dialect A and Lampung Saibatin Dialect O [1]. Unfortunately, Lampung is now classified as an endangered language as the number of speakers continues to decline. This decline is caused by the increasingly dominant use of Indonesian in daily communication, which slowly replaces the use of regional languages, including Lampung. Among the existing dialects, dialect A of Lampung Pepadun is still very


S. Hadi et al. (eds.), *Proceedings of the 5th International Conference on Applied Sciences, Mathematics, and Informatics (ICASMI 2024)*, Advances in Physics Research 13,
https://doi.org/10.2991/978-94-6463-730-4_10

poorly documented, which further emphasises the importance of preservation efforts. The preservation of Lampung language is crucial, not only to maintain linguistic and cultural diversity, but also as a form of respect for regional identity [2].

Technological developments in speech processing offer promising solutions to support language preservation efforts. One such technology is Speech-to-Text (STT) systems, which enable the automatic conversion of spoken language into text [3]. This technology can be applied in various fields, such as automatic translation, education, and language documentation. In the context of education, Speech-to-Text systems play an important role in supporting language learning and improving accessibility [4]. This technology is able to automatically transcribe spoken lessons into text in real time, which can improve students' comprehension and recall-especially in learning regional languages such as Lampung. In addition, this technology is very beneficial for students with hearing impairments or those who need visual assistance in the learning process. Educators can also utilise this technology to document and preserve oral traditions, such as folklore or oral literature, by converting them into written material for academic purposes [5].

The performance of Speech-to-Text systems depends on several important stages, including speech signal pre-processing, feature extraction, and the application of acoustic and language models to produce accurate transcriptions [6]. One of the widely used methods in speech feature extraction is Perceptual Linear Prediction (PLP), which is designed to mimic the way the human auditory system processes sound [7]. PLP's ability to adapt sound processing to human auditory perception makes it highly effective in dealing with languages that have complex phonetic structures [7]. Although PLP has been successfully applied to various languages, its use in Lampung Pepadun, particularly dialect A, is still under-researched. This research aims to examine how the PLP method can be used to capture the distinctive acoustic features of Lampung and convert them into accurate text representations.

With the decreasing number of native speakers, an effective Speech-to-Text system for Lampung language can be a very important tool in documenting daily conversations, traditional stories, as well as educational materials that could potentially be lost if not immediately recorded and preserved [8]. Given that language is both a means of communication and a store of cultural identity, the development and implementation of a successful Speech-to-Text system for Lampung language can be an important step in the effort to preserve endangered languages [9].

## 2 Methodology

This study uses Speech-to-Text (STT), approach based on Perceptual Linear Prediction (PLP) method for speech-to-text transcription in Lampung Pepadun language dialect A [10]. This methodology involves several main stages, including data collection, preprocessing, Perceptual Linear Prediction (PLP) feature extraction, and performance evaluation. In detail, the stages of this methodology are described as follows [7].

### 2.1 Data Collection

The data used in this study are voice recordings of native speakers of Lampung dialect A. The data collection process involves the participation of 4 native speakers who live in the Pepadun region spreading over Pesawaran Regency. Each speaker was asked to read the prepared sentences. Each voice recording was done with a high quality microphone and saved in WAV format. The audio was recorded at a sampling frequency of 16 kHz. Each speaker uttered the sentence 10 times, the total number of sentences recorded from all respondents amounted to 800 sentences. Table 1 shows 20 simple sentences that we used as a primary data which contains subject, predicate, object and adverb in Lampung language.

**Table 1.** lampung language sentences spoken by the speaker.

| No. | Lampung Sentence | English Translation |
|---|---|---|
| 1 | Ana ngebeli kupi di warung | Ana bought coffee at the stall |
| 2 | Daing main bal di lapangan | Brother plays soccer on the field |
| 3 | Lita mupoh pakaian di duwai | Auntie washes clothes in the river |
| 4 | Buyah ngunut iwa di duwai makai jaring | Father fishes in the river with a net |
| 5 | Sikam ngebaca buku di ruang tamu | We read books in the living room |
| 6 | Dede nanom sayuran di huma | Dede grows vegetables in the garden |
| 7 | Dia ngeguai kan guring di dapor | She cooks fried rice in the kitchen |
| 8 | Adek ngisik kucing di mahan | My sister has a cat at home |
| 9 | Ibuk ngelajar sanak ngebaca di kamar | Mom teaches the child to read in the room |
| 10 | Sikam nyiram bunga di pengadapan | I watered the flowers in the yard |
| 11 | Ama nanom pari di sabah | Uncle planted rice in the field |
| 12 | Atu ngebabay umpu di pengadapan | Grandfather holding grandson in the yard |
| 13 | Kucing pedom di lammung kerasi | Cat sleeping on a chair |
| 14 | Ateh ngusung keranjang di pasar | Grandma carries a basket at the market |
| 15 | Mak wan ngeberasihko mahan jeno pagi | Auntie cleans the house in the morning |
| 16 | Ateh ngayam sulan di debah batang | Grandmother weaving a mat under a trunk |
| 17 | Ayah ngusung mubil baru | Father drives a new car |
| 18 | Umik mupoh pakaian kamak | Mom washes dirty clothes |
| 19 | Adek nganik buah manga | Sister ate a mango |
| 20 | Daing main sepida di lapangan. | Brother playing bicycle on the field |

### 2.2 Data Pre-processing

Once the data is collected, the first step is to perform pre-processing to clean and prepare the data. This stage includes [11]:

- **Volume Normalisation**: Equalising the volume levels between recordings to reduce differences due to speaker voice variations.
- **Noise Removal**: Using the subtractive spectral method to reduce background noise that may appear during the recording process.

- **Segmentation**: The speech signal is divided into short 20ms frames with 50% overlap to capture subtle frequency changes.

### 2.3 PLP Feature Extraction

The main stage in this research is the feature extraction process using the Perceptual Linear Prediction (PLP) method. PLP is a feature extraction technique that mimics the way the human ear hears sounds. The PLP method is carried out through the following steps [7]:

- **Pre-Processing**: Before the voice signal is extracted, pre-processing is performed to clean the signal. This step includes volume normalisation to equalise the sound level and noise removal to reduce interference. The purpose of pre-processing is to ensure the signal quality is good enough for extraction.
- **Fast Fourier Transform (FFT)**: The next step is to convert the signal from the time domain to the frequency domain using the FFT. This process results in a spectral representation of the sound signal, which enables analysis of the frequencies in the sound signal.
- **Pre-Emphasis**: At this stage, the high frequencies are amplified to balance the signal intensity based on human perception, so that the high frequency details are clearer. This helps to highlight the changes in sound intensity that can be seen in the graph.
- **Bark Bank Filter**: The signal that has been converted to the frequency domain is passed through the Bark Bank Filter. This filter divides the frequency spectrum into bands that correspond to human frequency sensitivity. This allows PLP to mimic the way the human ear processes sound, prioritising the most important frequencies.
- **Logarithmic Energy Filter Bank**: At this stage, the energy of each frequency band is converted to a logarithmic scale. This compression allows the response to sound intensity to be non-linear, in keeping with the nature of human perception, so that the subtler changes in energy become more prominent.
- **Linear Prediction Coding (LPC)**: After going through the auditory filter, the signal is simplified using Linear Predictive Coding (LPC), which generates PLP coefficients. This technique creates a compact representation of the sound, which is used as the main feature in the PLP for further analysis.

In Perceptual Linear Prediction (PLP) feature extraction, there are 3 coefficients, which are blue, orange and green [7].

- **Coefficient 1 blue** is one of the coefficients generated from the Perceptual Linear Prediction (PLP) extraction process, which captures the dominant frequency component of the speech signal, this coefficient relates to the main spectral energy of the signal, reflecting the changes in intensity and frequency that occur during pronunciation. Dominant frequency here means the frequency component that has the greatest strength among other frequency bands.
- **Coefficient 2 orange** shows how the intermediate frequencies in the speech signal vary over time. These intermediate frequencies sit between the dominant components captured by Perceptual Linear Prediction (PLP) Coefficient 1 and the more

stable low frequencies, such as those captured by Perceptual Linear Prediction (PLP) coefficient 3. This coefficient reflects subtle changes in the speech signal, such as inter-phoneme transitions, intonation variations, or changes in speaking speed, which are important for understanding the pattern of reading by speakers.

- C**oefficient 3 green** captures the low frequency components of the speech signal, which are typically more stable than the dominant or mid frequencies. This coefficient provides information about the stability and less variable parts of the speech signal, such as when speakers utter long vowels or consonants that do not require much energy change, these low frequencies generally correspond to the more continuous components of the sound, which is important for understanding the basic sound patterns in speech.

The Perceptual Linear Prediction (PLP) extraction results from the speech signal are analyzed to ensure that the resulting features reflect the corresponding acoustic characteristics. Each Perceptual Linear Prediction (PLP) coefficient is analyzed based on the spectral variations recorded in the speech signal. The Perceptual Linear Prediction (PLP) graph displays the variation of the three main coefficients generated from the extraction process, showing how the PLP is able to capture important information from the sound signal [7].

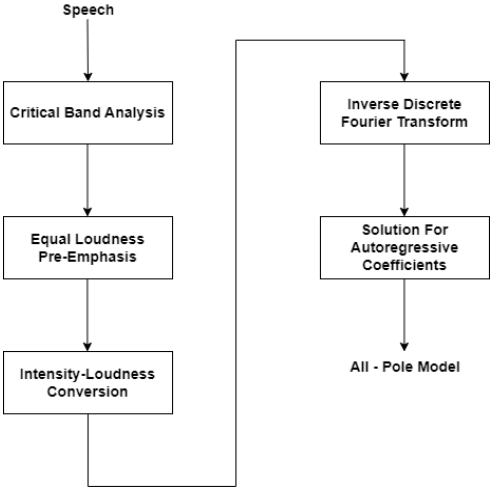


**Fig. 1.** Block Diagram of Perceptual Linear Prediction [7].

Figure 1 is a block diagram of Perceptual Linear Prediction (PLP). This diagram shows the Perceptual Linear Prediction (PLP) feature extraction process, which aims to generate PLP coefficients by imitating human auditory perception. PLP feature extraction produces feature vectors that will be used in the training and testing process of the STT model, this process improves speech representation by combining perceptual aspects such as critical band analysis, equal loudness pre-emphasis, and intensity-loudness conversion [7].

## 3 Results

The results of the Perceptual Linear Prediction (PLP) extraction show that the method successfully captures important spectral variations of the Lampung Pepadun Dialect A speech signal. The figure 2, 3, 4 and 5, shows a graph which generated from a file of 4 different speakers speaking the same sentence, sentence2.wav (Daing main bal di lapangan), shows three different Perceptual Linear Prediction (PLP) coefficients, each representing a different frequency characteristic of the speech signal.

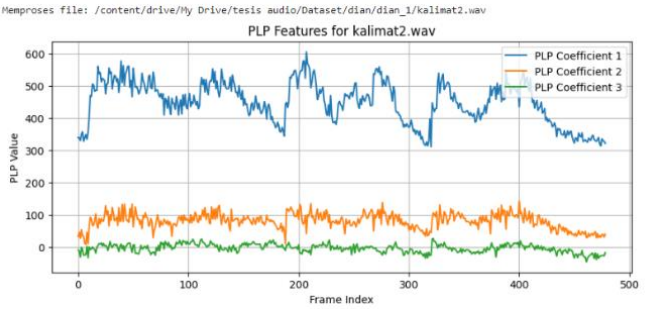


**Fig. 2.** PLP Extraction result of Dian Speaker.

- **PLP Coefficient 1 (Blue)**: This first coefficient of the first speaker shows significant fluctuations between frames 0 to 400, with peaks reaching values around 600. The largest fluctuations occur in the early frames (around frames 50-150), indicating that this speaker has large dominant frequency changes in the beginning of speech. After frame 200, the amplitude decreases gradually, showing stability in the latter part of the speech.
- **PLP Coefficient 2 (Orange)**: The second coefficient is relatively more stable, with less fluctuation, ranging from 0 to 100. It indicates smooth mid-frequency transitions throughout the speech signal.
- **PLP Coefficient 3 (Green)**: The third coefficient shows very little fluctuation, with values between -50 to 50, indicating that this first speaker has stability in the low frequency component during pronunciation.

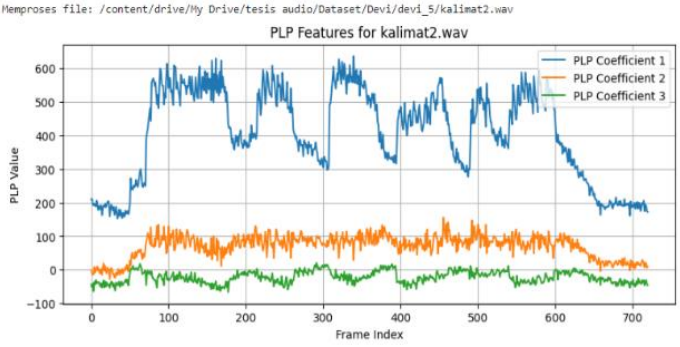

**Fig. 3.** PLP Extraction result of Devi Speaker.

- **PLP Coefficient 1 (Blue)**: The second speaker shows a more uniform pattern with several peaks throughout the signal, especially around frames 100 to 500. These peaks indicate strong changes in voice intensity throughout the spoken sentence, with more evenly distributed fluctuations compared to the first speaker.
- **PLP Coefficient 2 (Orange)**: The second coefficient shows slightly more dynamic fluctuations compared to the first speaker, with a wider range between 0 to 150, indicating more complex transitions in the mid-frequency part of the voice.
- **PLP Coefficient 3 (Green)**: The third coefficient shows a stable pattern, with a fixed value between -50 to 50, similar to the first speaker, indicating consistency in the low frequency component.

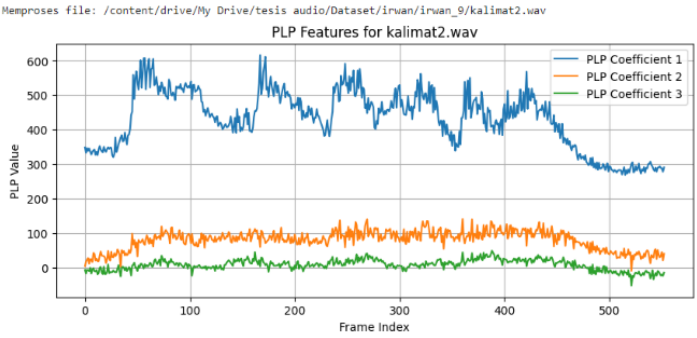


**Fig. 4.** PLP Extraction result of Irwan Speaker.

- **PLP Coefficient 1 (Blue)**: The third speaker shows a slightly different pattern, with large fluctuations around frames 100-500. The Perceptual Linear Prediction (PLP) Coefficient 1 values are over a very wide range, with peaks reaching over 600 in some parts. This indicates very strong frequency changes during pronunciation, especially in the middle and end of sentences.
- **PLP Coefficient 2 (Orange)**: The second coefficient shows more significant variation than the previous two, ranging from 0 to over 150. This indicates that these speakers have larger changes in the mid-frequencies, which could relate to accent or intonation.
- **PLP Coefficient 3 (Green)**: The third coefficient remained stable with values ranging from -50 to 50, indicating that the low frequency component in these speakers did not change much throughout the pronunciation.

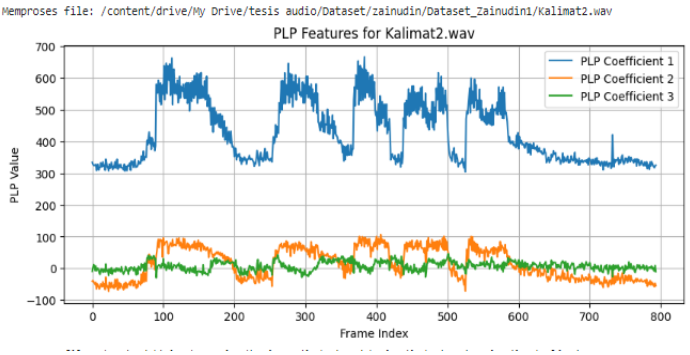


**Fig. 5.** PLP Extraction result of Zainudin Speaker.

- **PLP Coefficient 1 (Blue)**: The fourth speaker has a pattern more similar to the first speaker, but with a more evenly distributed amplitude throughout the signal. There are fluctuations in the beginning (frames 0-100) and middle (frames 300-400), indicating changes in the intensity of the voice at certain parts. The coffient blue shows that Zainudin has a stronger voice intensity or more significant emphasis in the way he pronounces sentences. This could be due to a more expressive speaking style or variations in word emphasis that are more pronounced.
- **PLP Coefficient 2 (Orange)**: The second coefficient shows a smoother pattern than the third speaker, with a smaller range (0 to 100), indicating simpler inter-phoneme transitions than the previous speaker. In coffient orange the orange line tends to be higher than the other three speakers. This indicates that Zainudin's voice has a more consistent intensity in the mid-frequency area, which may be due to stronger intonation or vocal resonance at those frequencies. This may reflect a speaking style that tends to have a deeper or more dominant intonation at mid-frequencies.
- **PLP Coefficient 3 (Green)**: The third coefficient is very stable throughout the signal, indicating a constant low frequency during pronunciation. Green's coffient is slightly higher than the other three speakers. This could indicate that the low frequency component in Zainudin's voice is more prominent or more stable. This could be due to his unique vocal structure or resonance, which makes his low frequencies more amplified.

## 4 Conclusion

Overall, this research has proven that the Perceptual Linear Prediction (PLP) method is a very effective approach to capture and analyse the acoustic features of the speech signal of Lampung Pepadun Dialect A. Perceptual Linear Prediction (PLP) is able to extract relevant speech characteristics.

Each PLP coefficient provides important information regarding the spoken sound element. Coefficient 1 (Blue), which captures the dominant frequency component, shows different intensity changes in each speaker, from large fluctuations at the begin-

ning (in Speaker 1) to an even amplitude distribution (in Speaker 4). Coefficient 2 (Orange) reflects mid-frequency variations that mark intonation transitions and different reading patterns; the speaker with more complex intonation (Speaker 3) has larger fluctuations. Meanwhile, coefficient 3 (Green) shows uniform low-frequency stability across speakers, although Speaker 4 shows a slight baseline rise that may be due to unique vocal structure or resonance.

With PLP's ability to recognise inter-speaker differences, detect dominant frequency components, and maintain stability at low frequencies, this method has great potential to be applied in local language-specific Speech-to-Text (STT) systems. This is particularly important in the context of local language preservation, where the documentation and processing of spoken languages is becoming increasingly urgent due to the diminishing native speakers.

Ultimately, this research makes an important contribution in building a technological foundation that can be used to support Lampung language preservation through a modern approach based on speech recognition. This research also opens up opportunities for further research that can expand the scope and improve the performance of speech recognition systems for other local languages.

## References

1. Amalia, U., Surya, A., Ardiansyah, M.: Revitalisasi Bahasa Lampung melalui Pendidikan dan Teknologi. Jurnal Pelestarian Bahasa Nusantara **8**(2), 15-25 (2023)
2. Jumlah Penutur Bahasa Lampung, https://lampung.antaranews.com/berita/706452/kantor-bahasa-sebut-jumlah-penutur-bahasa-lampung, last accessed 2024/08/25
3. Graves, A., Jaitly, N., Mohamed, A.: Hybrid Speech Recognition with Deep Bidirectional LSTM. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 273-277. (2013)
4. Jurafsky, D., J. Martin, H.: Speech and Language Processing. 3rd edn. Prentice Hall (2021)
5. Kartika, N., Dienaputra, R. D., Machdalena, S., Nugraha, A., Sriwardani, N.: Oral Tradition in Preserving the Natural Environment in Kampung Adat Dukuh, Ciroyom Village, Cikelet Subdistrict, Garut District. Mudra Jurnal Seni Budaya **37**(3), 247–256 (2022)
6. Bahdanau, D., Cho, K., Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. (2014)
7. Hermansky, H.: Perceptual Linear Predictive (PLP) Analysis of Speech. The Journal of the Acoustical Society of America **87**(4), 1738-1752 (1990)
8. Bishop, C. M.: Pattern Recognition and Machine Learning. Springer, (2006)
9. Huang, X., Acero, A., Hon, H.-W.: Spoken Language Processing: A Guide to Theory, Algorithms, and System Development. 1st edn. Prentice Hall, USA (2001)
10. Povey, D. et al.: The Kaldi Speech Recognition Toolkit. IEEE ASRU (2011)
11. Jelinek, F.: Statistical Methods for Speech Recognition. MIT Press (1997)