தமிழில் பேச்சுத் தொழில்நுட்பம் (Speech Technology and Tamil)

ஆ. க. ராமகிருட்டிணன் (A G Ramakrishnan)

மனுமொழி ஆய்வகம், மின் பொறியியல் துறை, இந்திய அறிவியல் பயிலகம், பெங்களூரு 560012.

MILE Laboratory, Dept. of Electrical Engineering, Indian Institute of Science, Bangalore 560012.

கட்டுரைச் சுருக்கம் (Abstract)

Development of text to speech and speech recognition technology in Tamil shall make interaction with the computing devices directly in Tamil eminently possible for the common man, even if he is not technically savvy. This will make equal access to knowledge for every Tamil citizen. While viable text to speech (TTS) technology in Tamil has already been developed by us based on unit selection synthesis, there is hardly any research or development work going on in large vocabulary automated speech recognition (ASR) in Tamil. The conventional speech recognition paradigms that have been used for the European languages is ill suited for Tamil, since Tamil is a morphologically very rich language. The only speech recognition technology that currently is reasonably successful that is also used in commercial ASR systems for European languages makes use of a huge vocabulary and is essentially a transcription system that finds the sequence of words from the fixed vocabulary that best matches the input speech signal. This model cannot be directly used for most of the Dravidian languages such as Tamil, since these languages possess an exponentially growing lexicon. While the current commercial English ASRs employ a lexicon of about two lakh words, our own work has identified over sixteen lakh unique words in Tamil and the size is still growing.

This article first gives the design and development details of our Thirukkural TTS system. Then, it proposes a new, alternate, phoneme recognition based approach to automated speech recognition that must be explored to develop workable recognition technologies for Tamil.

Keywords: speech synthesis, text normalization, morphology, Tamil, Dravidian languages, automated speech recognition, bigram models, domain dependent lexicons, naturalness, intelligibility, unique words

Thirukkural (திருக்குரல்) Speech Synthesis System - Introduction

We have developed a good text-to-speech (TTS) conversion system for Tamil language [1] and the web demo is available on the MILE lab website of Indian Institute of Science, Bangalore [2]. It involves two phases, namely, the offline phase and the online phase [3].

Offline phase includes collection of a phonetically rich synthesis speech database and segmenting and pitch marking the same. Online phase includes text analysis, waveform concatenative synthesis after unit selection and signal processing for prosody. The different methods for synthesizing speech are first reviewed. Then, the offline processes are explained, followed by the online phase. Subsequently, the implementation details are discussed and finally, the potential applications.

The techniques employed for synthesizing speech from text may be broadly classified into three categories:

I. Articulatory synthesis: In the articulatory synthesis approach, each phone is modelled by a huge number of parameters that describe the relative positions of the articulators in speech production. From detailed static and dynamic articulatory measurements, the movement is parameterized and used for the synthesis of speech sounds. Though it is useful for a study of speech production mechanisms, so far it has not produced a very natural quality speech [4].

II. Model-based synthesis: Here, the waveforms are modelled using linear prediction (LP) coefficients. The linear prediction model is an all-pole model which models vowels exceptionally well, but fails to model the nasals and silence (stops).

III. Concatenation-based synthesis: In this approach, segments of natural speech are concatenated to form the speech output corresponding to the input text. This is more natural than the above techniques, but the size of the speech database required is fairly large. In concatenation with waveform modification, speech from arbitrary text can be synthesized with reasonable quality. These systems have flexibility in selecting the speech segments to concatenate because the waveforms can be modified to allow for a better prosody match. This means that the number of sentences with mediocre quality is lower than the case where no prosody modification is allowed. On the other hand, replacing natural with synthetic prosody can hurt the overall quality. In addition, the prosody modification process also degrades the overall quality. Figure 1 shows the block diagram of our TTS system using concatenation principles, including waveform modification [5].

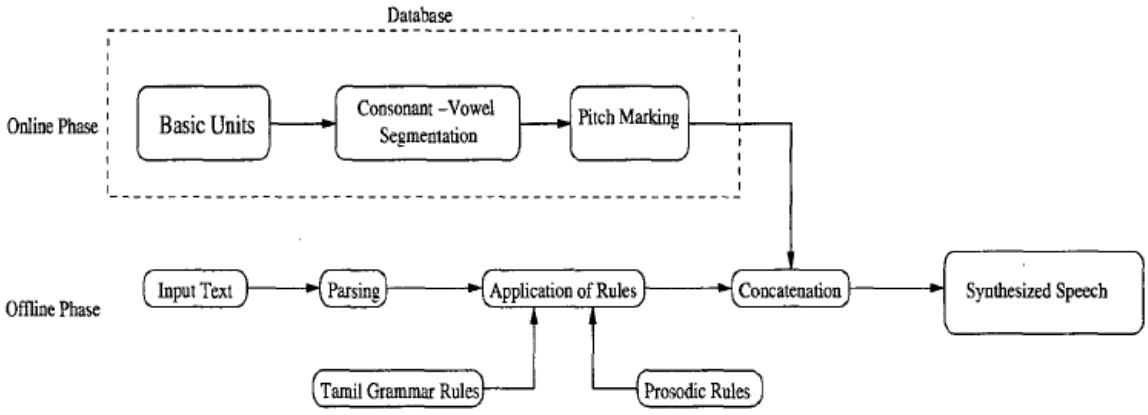


Fig. 1. Block diagram of the Thirukkural Tamil speech synthesis system.

Offline and online processes

The offline processes of the system include (1) Choosing the basic units (2) Building the database (3) Detailed study of prosody in natural speech and modelling (4) Consonant - vowel segmentation (5) Pitch marking. The different online modules that need to be developed are text normalization, grapheme to phoneme conversion, prosody prediction, dividing the phonetic text into the basic units, unit selection, concatenation and modification of pitch, duration and amplitude based on the prosody model.

Choosing the basic units

The fundamental building block of speech is a phoneme. The basic units that can be used for synthesis are diphones, triphones, demi-syllables, syllables, words, phrases and sentences. In terms of synthesis quality, sentence is the best choice for basic unit and phoneme is the worst. However, the size of the database is also an important factor to be considered. In order to minimize the distortion due to concatenation, we must have fewer concatenations and thus use long units such as words, phrases or even sentences. However, practically infinite units must be stored if the basic units are sentences.

Whereas it is all right to have units with prosody slightly different from the desired target, replacing a unit having a rising pitch with another having a falling pitch will result in an unnatural sentence. The units must be of a general nature, if conversion of any unrestricted Tamil text is desired. However, if we choose words or phrases as our units, we cannot synthesize arbitrary speech from text, because it is almost guaranteed that the text will contain words not in our inventory. The use of arbitrarily long units results in low concatenation distortion but our study of a huge Tamil text corpus shows that the number of unique words in Tamil far exceeds 16 lakh [6]. The longer the speech segments are, the more of them we need, to be able to synthesize speech from arbitrary text.

Considering the above issues, syllables or demisyllables have been used as basic units. This may contain phonemes, diphones or triphones. The synthesis database was recorded from a native Tamil speaker. Recording took place in a noise free room using Shure SM 58 microphone, whose frequency response is 50 - 15,000 Hz, connected to a PC. Carefully chosen sentences were recorded at a high sampling rate.

Modeling prosody in natural speech

Prosody is a complex weave of physical, phonetic effects that is being employed for expression. Prosody consists of systematic perception and recovery of the speaker's intentions based on pauses, pitch, duration and loudness. Pauses are used to indicate phrase boundaries and to indicate the end of a sentence or a paragraph. It has been observed that the silence in speech increases as we go from comma to end of sentence to end of paragraph. Pitch is the most expressive part of a speech signal.

Need for pitch modification

We try to express our emotion through pitch variation. Constant pitch signal sounds very unnatural. Figure 2 illustrates how the pitch and amplitude of the signal varies across a sentence. If a sentence is rendered as a declarative utterance, the amplitude and the pitch slowly droop down in the last word. When the same sentence is spoken as an interrogative utterance, both the pitch and the amplitude contour increase last word.

We have proposed two methods to estimate the local pitch values. One method uses the fact that the discrete cosine transform [DCT] of the speech signal displays prominent harmonics of the pitch frequency [7]. The second approach uses DCT based spectral autocorrelation [8]. Pitch marking is essential as the waveforms are concatenated at the pitch marks. The method employed for pitch marking is based on the integrated linear prediction residual (ILPR). The dynamic plosion index (DPI) of half-wave rectified ILPR is computed at the current pitch mark and the next epoch location is identified using the maximum peak to valley swing in the DPI sequence in the immediate 15 msec interval [9]. Wherever needed, the pitch is modified by a pitch synchronous pitch scaling technique proposed by us [10]. Here, the signal corresponding to a pitch period is inverse filtered to obtain the LP residual, whose length is modified by truncation or zero padding of its DCT sequence. The length modified LP residual is passed through the forward linear prediction filter to obtain the pitch modified speech.

Duration is the second important factor that affects the naturalness of the synthesized speech. Same vowel appearing in different positions in a word or a sentence has different durations. For example, consider the sentence "நான் ஆறு மணிக்கு வரலாமா?". In this sentence, vowel /ஆ/ appears at different positions as shown in Table 1. Similarly, Table 2 shows how the duration of the word "சென்றான்" varies widely depending upon its position in and the length of the sentence.

அட்டவணை 1: சொல்லில் எழுத்தின் இடத்தைப் பொறுத்து கால அளவின் மாற்றம்.

சொல்	சொல்லில் 'ஆ'வின் இடம்	'ஆ'வின் கால அளவு (நிமிடம்)
நான்	நடுவில்	0.15
ஆறு	முதலில்	0.16
வரலாமா	நடுவில்	0.15
வரலாமா	கடைசியில்	0.18

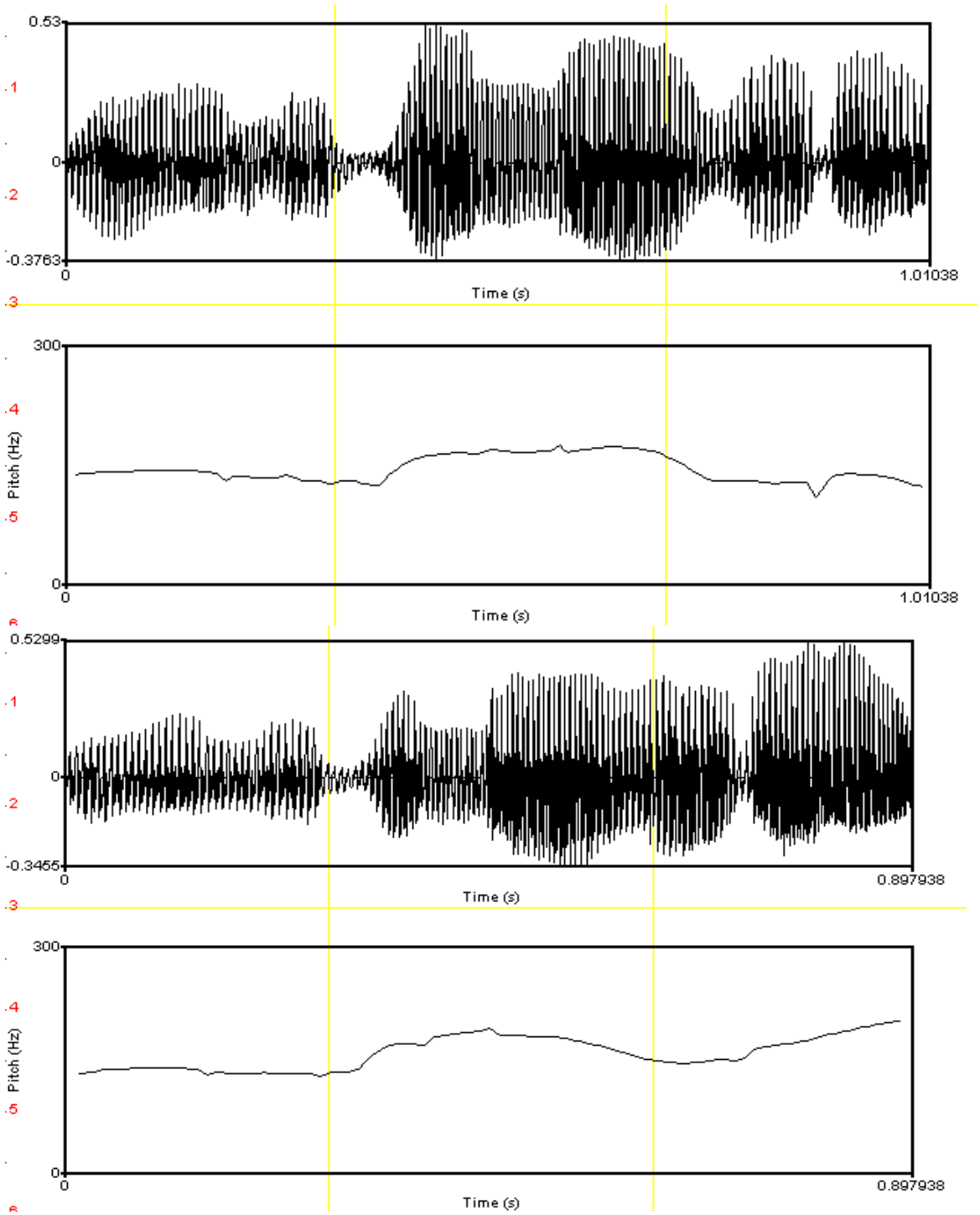


Fig. 2. Prosody for naturalness. The first is the affirmative utterance of the sentence "அவன் நெஜம்மா வரான்."; the pitch and the amplitude of the waveform droop down towards the end of the utterance. The second is the same sentence uttered as a question; here, pitch and amplitude increase at the end.

அட்டவணை 2: ஒரு சொல்லின் கால அளவு வாக்கியத்தில் அதன் இடத்தைப் பொறுத்து மாறுபடுகிறது.

Utterance	Duration of "சென்றான்"
சென்றான்.	820 mSec
அவன் சென்றான்.	687 mSec
அவனும் சென்றான், அவளும் சென்றாள்.	631 mSec
அவன் சென்றான் எனக் கேள்விப்பட்டேன்.	489 mSec

Consonant Vowel Segmentation

It is observed that any change in the consonant part of a signal results in change of perception of the unit. Consonants must be kept intact. To this end, consonant and the vowel regions of the units must be segmented. A consonant may or may not be co-articulated with the preceding or following vowel. Non co-articulated consonant can be segmented easily using difference of energy in consecutive blocks (frames of short duration) of the signal. The given speech unit is divided into frames of 10 msec duration each. Energy of each frame is calculated and the first difference of the energy contour gives two distinct peaks, one on the positive side and the other, on the negative side. Co-articulated consonant segmentation is more challenging than the other. Energy contour is almost flat at the transition from vowel to consonant. We have proposed a method of segmentation of co-articulated consonants (semi-vowels) from the adjoining vowels using projection into consonant and vowel subspaces [11], which gives much better segmentation. However, for better accuracy, we resorted to manual correction. The results of segmentation of a sample non-co-articulated and co-articulated consonant each are shown in Fig. 3.

Online Process

This involves two important phases: (a) Text analysis (b) Synthesis. Text analysis phase involves first text normalization, which converts all the abbreviations, numerals and other symbolic notations into the corresponding Tamil text strings [12]. Then, it converts the input text into a phoneme sequence by the application of Tamil grapheme to phoneme rules [13] and then parses it into a sequence of basic units of speech. Synthesis phase involves the concatenation of the waveforms of these units in the correct sequence using the pitch continuity metric [14] and synthesis after application of pitch modification [10], if necessary.

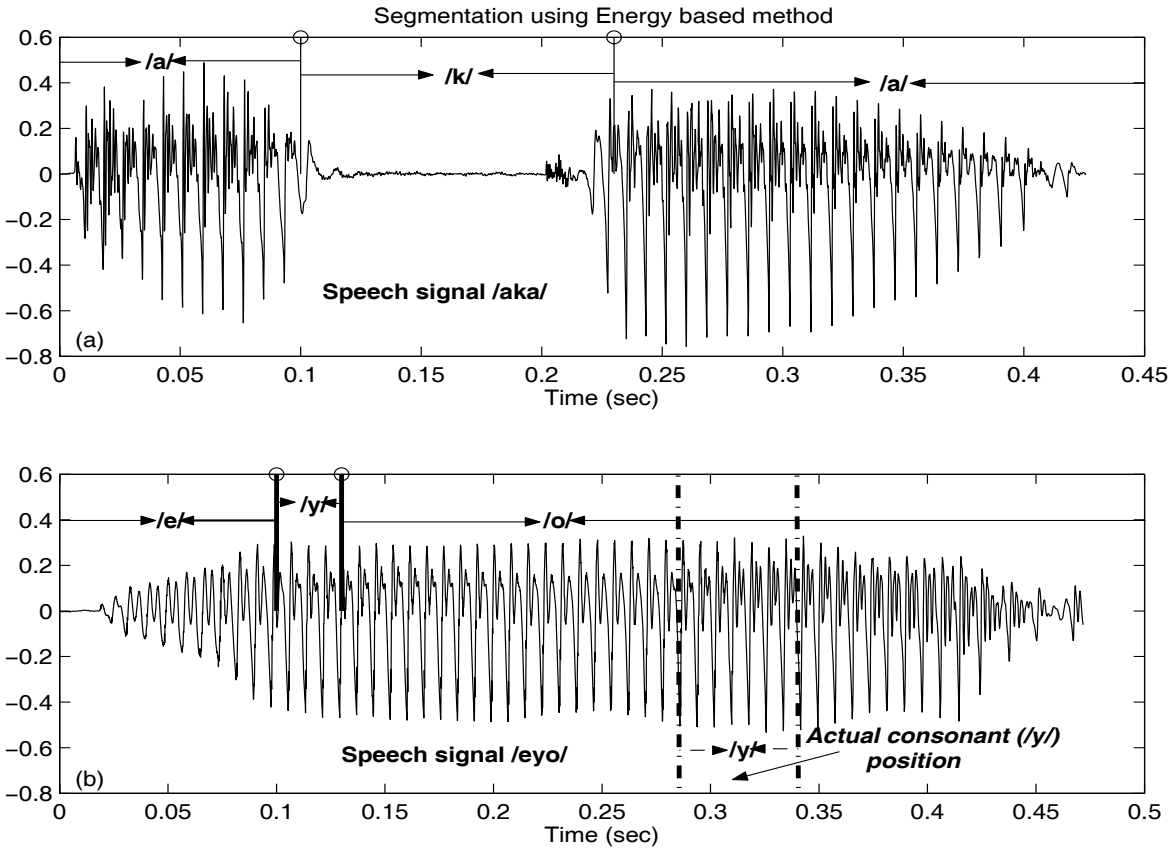


Fig. 3. Top: Segmented and pitch-marked non-coarticulated VCV / அக/. Bottom: coarticulated VCV /இயி/, where boundaries cannot be easily detected using energy differences.

Implementation

We tried to use the Festival speech synthesis framework [15] but after considerable work, decided that we needed to develop our own framework for text to speech synthesis [5], so that various research experiments can be carried out easily. So, the system is designed and developed from scratch using our own algorithms and code, without borrowing any code or framework from any outside source. It works on Windows and Linux. The output is stored in wave file format. It takes Unicode text as input and the input can be keyed in using our own Indic Keyboard IME, which is a open source software that can be downloaded from code.google.com [16].

Applications of TTS

975 visually challenged members from all over Tamilnadu are regularly getting short information from Anna Centenary Library in Chennai converted into speech using our web TTS demo. Thirukkural TTS system has a male voice and can read input text in Unicode format and synthesize intelligible and acceptably natural Tamil speech [17]. That the quality of speech is good is proved by the second place in the International Blizzard TTS Challenge

2013 for Indian languages conducted by University of Edinburgh [18]. Attempts are being made to make it natural [19], add emotions [20. We have also simulated certain experiments to explore whether the TTS can be ported on to a mobile platform [21]. The other applications of speech synthesis in Tamil are: (1) Natural language interface for computers; (2) Self-learning multimedia education packages in Tamil; (3) Automatic telephone-based enquiry systems in all Government organisations; (4) Audio on-line help in all Tamil language based IT software; (5) Computer or web based Tamil teaching; (6) Automatic document reading machines in Tamil for the blind; (7) To synthesize different types of voices in animation movies; (8) Digital personal assistant (English/Indian language to Tamil).

உணர்விகளில் தமிழ் மொழி மாதிரியங்கள்

இந்தக் கட்டுரையின் இனி வரும் பகுதியானது தமிழில் வாய்மொழி உணர்தல் (speech recognition) முறைமைகளில் மொழி மாதிரியங்களின் (language models) பயன்பாட்டில் எதிர்கொள்ளப்படும் சில குறிப்பிட்ட இடர்பாடுகளை விவாதிக்கிறது. இத்தகைய மொழி மாதிரியங்கள் ஐரோப்பிய மொழிகளில் மிகுந்த இலாபகரமாகப் பயன்படுத்தப்பட்டாலும், அவை அவ்வாறே தமிழில் பயன்படுத்தப்படும்பொழுது, தமிழின் மொழியியல் செறிவு காரணமாக அம்மொழிகளில் காணப்படாத சில புதிய இடர்பாடுகளை எதிர்கொள்ள நேரிடுகிறது. இது தமிழின் புணர்ச்சி விதிகள் (sandhi rules) மற்றும் உருபனியல் மாற்றச் செறிவு (morphological richness) காரணமாகவும் சொற்களோடு விகுதிகள் இணைந்து பொருள்தரும் பண்பு (agglutination) மற்றும் பகுதி கட்டில்லாச் சொல்வரிசை முறை (Partially free word order) காரணமாகவும் ஏற்படுகிறது. ஒரு மீப்பெரு பனுவலை (large corpus) அல்லது பனுவலில் இருந்து பெறப்பட்ட பெரும் சொல்வங்கியை n-கிராம் மொழி மாதிரியம் கொண்டு பகுப்பாய்வு செய்ததன் மூலமாகப் பெறப்பட்ட பல்வேறு புள்ளி விவரங்களை முன் வைத்து, அதன் மூலம் தமிழில் எத்தகைய மொழி மாதிரியங்களை எவ்வாறு பயன்படுத்தலாம் என்பது குறித்த என் கருத்துக்களையும் இங்கு சமர்ப்பிக்கிறேன்.

தமிழில் வாய்மொழி உணர்தல்

தற்போதைய கணினிமுறை பெருஞ் சொற்களஞ்சிய தொடர் வாய்மொழி உணர்தல் (large vocabulary continuous speech recognition - LVCSR) முறைமைகள் 2,00,000-க்கும் மேற்பட்ட எண்ணிக்கையிலான பெரும் சொல்லகராதியைப் பயன்படுத்துகின்றன. சொற்களஞ்சியத்தின் பருமன் அதிகரிக்கும்போது அதையொத்து விட்டர்பி சொல்

அணிக்கோவை (word lattice), இன்ன பிற தேடுதல்களில் கணிப்பு சிக்கற்பாடு அதிகரிக்கின்றது. மொழியியல் செறிவு காரணமாகத் தமிழ் சொல்வங்கியானது பனுவலின் அதிகரிப்பு விகிதத்துக்கு ஏற்ப (மடக்கை விகிதத்தில்) அதிகரிக்கிறது. ஒரு வினை வேர்ச்சொல்லின் (root verb) வருவித்த படிமங்கள் (derived forms) மிகச் சிலவே உள்ள ஆங்கிலம் மற்றும் ஹிந்தி போன்ற மொழிகளைப் போலல்லாமல் தமிழானது ஒரு வேர்ச்சொல்லுக்குச் சில ஆயிரங்களுக்கு மேற்பட்ட வருவித்த படிமங்களைக் கொண்டுள்ளது [22]. வினைச்சொற்களின் ஒவ்வொரு உருபனியல் கூட்டப்பட்ட வடிவமும் காலம், இடம், பொருள், ஏவல், மறுப்பு, அழுத்தம், கேள்வி போன்ற பல்வேறு வகையான பொருளை வெளிப்படுத்தும் இயல்பில் புதிய எல்லைகளைக் கொண்டுள்ளது. சொல்லிலக்கண விதிகளில் அருகாமை சொல்லிலக்கணக் கூறுகளுடன் இணையும் தன்மையானது ஒருவிதமான நெகிழ்வுத் தன்மையைக் கொண்டுள்ளது. இவ்விதமான நெகிழ்வுத் தன்மையானது சொல்லிலக்கணக் கூறுகள் வெவ்வேறு இடங்களில் இணைய அனுமதிக்கிறது. இதனால் தமிழில் உருபனியல் மாற்றச் செறிவானது வினைச்சொற்களின் பல்வேறு வடிவங்களின் எண்ணிக்கையை ஆச்சரியப்படத்தக்க வகையில் அதிகரிக்கிறது [23]. உதாரணமாக நாங்கள் பயன்படுத்தும் வரையறுக்கப்பட்ட சொல்வங்கியில் உள்ள 'வா' என்கிற வேர்ச்சொல்லின் வருவித்த படிவங்களின் தனிப்பட்ட எண்ணிக்கை 4567 ஆகும். இதன் மூலமாகத் தமிழ் போன்ற செறிந்த மொழிகளுக்கான உணர்தல் மூலோபாயங்கள் (recognition strategies) ஆங்கிலம் மற்றும் ஹிந்தி போன்ற மொழிகளுக்கான மூலோபாயங்களை ஒத்திருக்கவியலாது என்பது புலனாகிறது. எனவே தமிழுக்கு ஏற்றாற்போல மாற்று மூலோபாயங்கள் வகுக்கப்பட வேண்டியதின் அவசியம் இங்கு வலியுறுத்தப்படுகிறது.

வாய்மொழி உணர்வியில் மொழி மாதிரியங்கள்

பெருஞ் சொற்களஞ்சிய தொடர் வாய்மொழி உணர்தல் (LVCSR) முறைமைகள் இரண்டு லட்சம் சொற்களுக்கும் மேற்பட்ட எண்ணிக்கையிலான பெரும் சொல்லகராதியைப் பயன்படுத்துகின்றன. ஆனால் ஒரு சராசரி மனிதனின் மூளையில் பதிந்திருக்கும் சொற்களின் எண்ணிக்கை அதாவது மனிதனின் கற்றுக்கொண்ட சொல்லகராதி இருபதாயிரத்திலிருந்து [24] ஐம்பதாயிரம் [25] வரை தான். இதனால்,

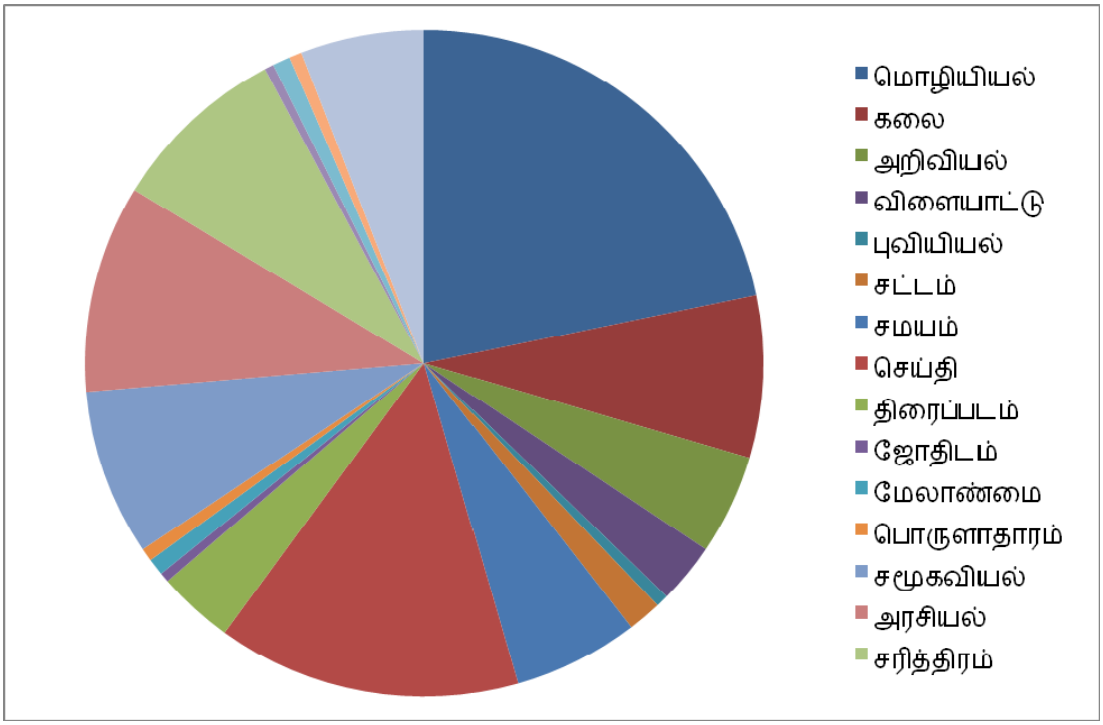
கணினிகளுக்கு மட்டுமே சாத்தியமாகக்கூடிய ஒரு மீப்பெரு கட்டற்ற சொல்லகராதியானது, உணர்தல் முறைமைகளில் மேலும் குழப்பங்களைத் தருவிக்கவே வாய்ப்பு அதிகம் என்பது அறியப்படுகிறது. நாம் உரையாடும்போது, ஆள்களம் நமக்கு பேசுவதற்கு முன்பே பிடிபட்டுவிடுகிறது. அதன்பின், பல்பொருள் ஒருமொழிச் சொற்கள் பயன்படுத்தப்படினும், மனித மூளையின் தேடுதலானது, அச்சமயத்தில் விவாதிக்கப்படும் ஆள்களத்துக்கு உட்பட்டே அமைகிறது. மேலும் கட்டயடுதல் செய்பணி (dictation tasks) முறைகள் பொதுவாக ஒரே அல்லது ஒருசில தொடர்புடைய ஆள்களங்களுக்கு உட்பட்டே அமைகின்றன.

இந்நிலையில் தகுந்த ஆள்களத்தை முன்னறிவதன் மூலமாக அறியப்படும் அறிவுப்புலத்தை வாய்மொழி மற்றும் எழுத்துரு கண்டறிதலில் பயன்படுத்துவது இத்தகைய குழப்பங்களைக் குறைப்பது மட்டுமில்லாது, கணிப்பு சிக்கற்பாட்டையும் குறைத்து துல்லிய அளவையின் படித்தரத்தை உயர்த்தும் என்ற கோட்பாட்டை இங்கு முன்மொழிகிறோம். இந்நோக்கில் இங்கு 2.2 கோடிக்கும் மேற்பட்ட சொற்களைக் கொண்ட பனுவல் பகுப்பாய்வுக்கு உட்படுத்தப்பட்டுள்ளது. இப்பனுவலானது 14 தனிப்பட்ட ஆள்களங்களாக (domains) வகைப்படுத்தப்பட்டுள்ளது. அவை முறையே மொழியியல், கலை, விளையாட்டு, புவியியல், சட்டம், சமயம், அறிவியல், செய்தி, அரசியல், திரைப்படம், ஜோதிடம், மேலாண்மை, பொருளாதாரம் மற்றும் சமூகவியல் ஆகியவை . இவற்றிலிருந்து பல்வகைப்பட்ட ஆள்களங்களில் உள்ள தனிச்சொற்கள், சொல் மற்றும் எழுத்து மட்ட n-கிராம் புள்ளிவிவரங்கள் போன்ற பலவகைப்பட்ட பகுப்பாய்வு முடிவுகள் தருவிக்கப்பட்டுள்ளன. அட்டவணை 3, சில பொதுவான ஆள்களங்களையும் அவை ஒவ்வொன்றிலும் உள்ள தனிச்சொற்களின் எண்ணிக்கையையும் அவற்றின் குறுக்க மற்றும் ஒப்பீட்டு விகிதங்களையும் காட்டுகிறது. உதாரணமாக, மொத்தச் சொல்வங்கியில் உள்ள தனிச்சொற்களின் எண்ணிக்கை 17.16 லட்சமாக இருந்தபோதிலும் 'அறிவியல்' ஆள்களத்தில் உள்ள தனிச்சொற்களின் எண்ணிக்கை 136931 ஆகும். ஆக, நாம் அறிவியல் துறை சார்ந்த ஒரு சொல்வதெழுதல் பணியில் ஈடுபடும்போது, 136931 சொற்களாலான சொல்வங்கியைப் பயன்படுத்தினால் போதுமானது. இது மென்பொருளின் வேகத்தையும் துல்லியத்தையும் நன்கு

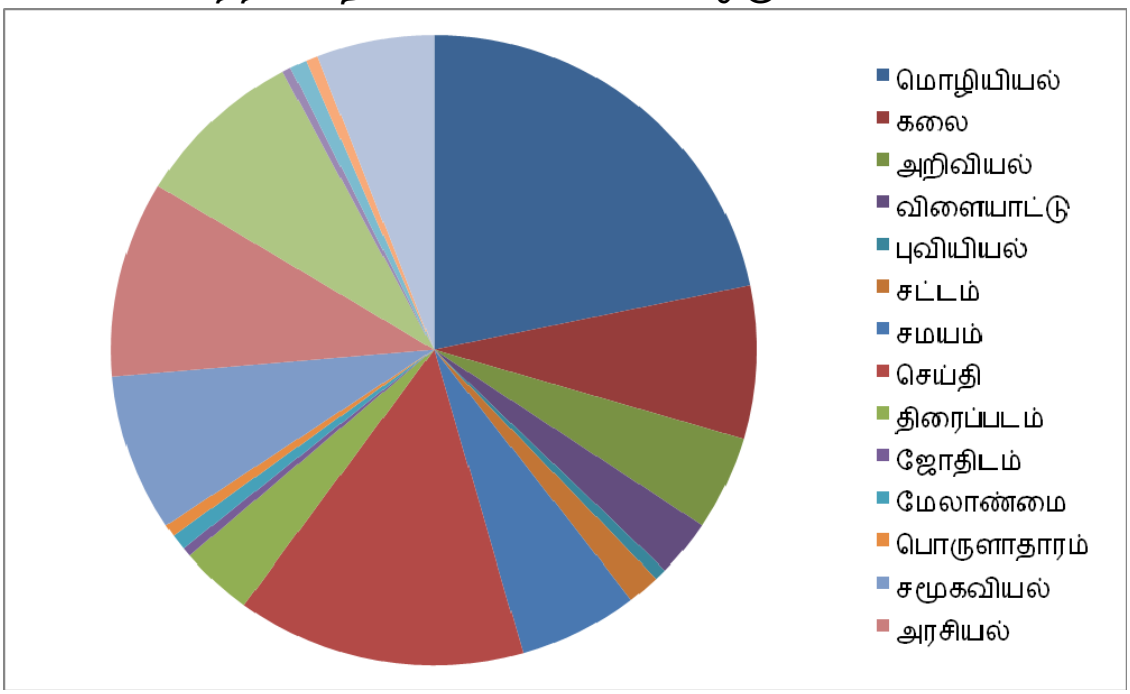
அதிகரிக்கும் என்பது திண்ணம். வட்டப் படங்கள் 4 மற்றும் 5-ம் மேற்கூறிய கருத்துக்களை வலியுறுத்துகின்றன.

அட்டவணை 3: தமிழ்ச் சொல்வங்கியின் அளவு, தனிச்சொற்களின் எண்ணிக்கை, ஆள்களத்தைத் தேர்வு செய்வதால் பெறப்படும் பேரகராதி மற்றும் சிக்கற்பாட்டு குறுக்கம், வரையறுக்கப்படாத ஒற்றைப் பேரகராதியின் தனிச்சொற்களின் எண்ணிக்கை : ஓர் ஒப்பீடு.

ஆள்களம்	சொற்களின் எண்ணிக்கை	தனிச்-சொற்களின் எண்ணிக்கை	சொல்லக-ராதியின் குறுக்க விகிதம் (%)	ஆள்களத்துக்கு உட்பட்ட ஒப்பீட்டு விகிதம் (%)
மொழியியல்	4179938	614184	35.8	85.3
கலை	1055330	222333	12.9	78.9
அறிவியல்	769531	136931	7.9	82.2
விளையாட்டு	831656	82384	4.8	90.0
புவியியல்	59124	17051	0.99	71.2
சட்டம்	189321	45738	2.7	75.8
சமயம்	878975	169277	9.8	80.7
செய்தி	6056139	408833	23.8	93.2
திரைப்படம்	779432	103163	6.0	86.7
ஜோதிடம்	37757	13194	0.7	65.1
மேலாண்மை	85183	23116	1.3	72.8
பொருளாதாரம்	120157	17873	1.0	85.1
சமூகவியல்	1309273	227952	13.3	82.6
அரசியல்	3779525	285040	16.6	92.5
சரித்திரம்	1681028	241082	14.0	85.6
வேளாண்மை	42754	12396	0.7	71.0
கல்வி	126546	23749	1.4	81.2
கவிதை	30105	17065	0.99	56.7
மற்ற பிற	699317	166283	9.7	76.2
மொத்தம்	22707568	1716256	--	92.5



படம் 4. மநுமொழி தமிழ் பனுவலில் ஆள்களம் வாரியாக மொத்த சொற்களின் எண்ணிக்கை - ஒரு வட்டப் படம்.



படம் 5. மநுமொழி தமிழ் பனுவலில் ஆள்களம் வாரியாக தனிச் சொற்களின் எண்ணிக்கை - ஒரு வட்டப் படம்.

எவ்விரு ஆள்களங்களுக்குள்ளும் பொதுவான சொற்கள் குறைவே.

நான்காம் அட்டவணை, ஒரு உதாரணமாக, "சமயம்" என்ற ஒரு ஆள்களத்துடன் வெவ்வேறு ஆள்களங்கள் தனித்தனியே சேரும்போது மொத்தச்சொற்கள் எத்தனை, அவை இரண்டிற்கும் பொதுவான சொற்கள் எத்தனை என்பதைக் காட்டுகிறது. "சமயத்"தில் தனிச்சொற்கள் 169277 உள்ளன. மொத்த சொற்களுடன் ஒப்பிடும்போது பொதுச் சொற்கள் 13 அல்லது அதற்கும் குறைவான விழுக்காடுகளே உள்ளன. ஆகையால் ஆள் கள ம் பொருத்தே நாம் சொல் வங்கியைப் பயன்படுத்த வேன்டும் . அட்டவணை 4 ஒரு மாதிரியே. உண்மையில் 19 ஆள்களங்களுக்கிடையே, 171 இருகள பிணைப்புகளைச் சோதிக்க முடியும்.

அட்டவணை 4: இரு ஆள்களங்களுக்கிடையே உள்ள பொதுவான தனிச்சொற்கள். எல்லா வரிசைகளுக்கும் பொதுவான ஆள்களமான "சமயம்", 169277 தனிச்சொற்கள் கொண்டுள்ளது.

ஆள்களம்	தனிச் சொற்கள்	மொத்தம் (கூட்டு)	பொதுவான சொற்கள்
திரைப்படம்	103163	260853	11587
கலை	222333	345710	45900
விளையாட்டு	82384	240884	10777

பல ஆள்களங்களுக்குப் பொதுவான தனிச்சொற்களின் எண்ணிக்கை

அட்டவணை 5-ஆனது ஆள்களங்களுக்குப் பொதுவான தனிச்சொற்களின் எண்ணிக்கை, ஆள்களங்கள் சேரச் சேர வெகு விரைவில் குறைந்து கொண்டே வருவதைக் காட்டுகிறது. இது, தனிச்சொற்களும் ஒவ்வொரு துறைக்கும் பிரத்தியேகமானவையே என்பதை உணர்த்துகிறது. ஆக எல்லாத்துறைகளையும் சேர்த்த மொத்தச்சொல் வங்கியை உபயோகிப்பதில் எவ்வித லாபமும் இல்லை என்பது புலனாகிறது.

அட்டவணை 5: இரண்டு, மூன்று, நான்கு எனப் பல ஆள்களங்களுக்குப் பொதுவான தனிச்சொற்கள். ஆள்களங்களின் எண்ணிக்கை அதிகமாக, அதிகமாக பொதுச் சொற்கள் குறைந்து கொண்டே வருகின்றன.

ஆள்களம்	தனிச் சொற்கள்	முழுவதற்கும் பொதுவான சொற்கள்
சமயம்	169277	169277
திரைப்படம்	103163	11587
கலை	222333	9960
விளையாட்டு	82384	5887
கவிதை	17065	1854
அறிவியல்	136931	1705
கல்வி	23749	1046
சரித்திரம்	241082	1044
புவியியல்	17051	711
சட்டம்	45738	680
மற்ற பிற	166283	680
மேலாண்மை	23116	589
ஜோதிடம்	13194	503
வேளாண்மை	12396	421
செய்தி	408833	420
பொருளாதாரம்	17873	392
அறிவியல்	227952	392
மொழியியல்	614184	392
அரசியல்	285040	392

கையெழுத்து உணர்வியில் மொழி மாதிரியங்கள்

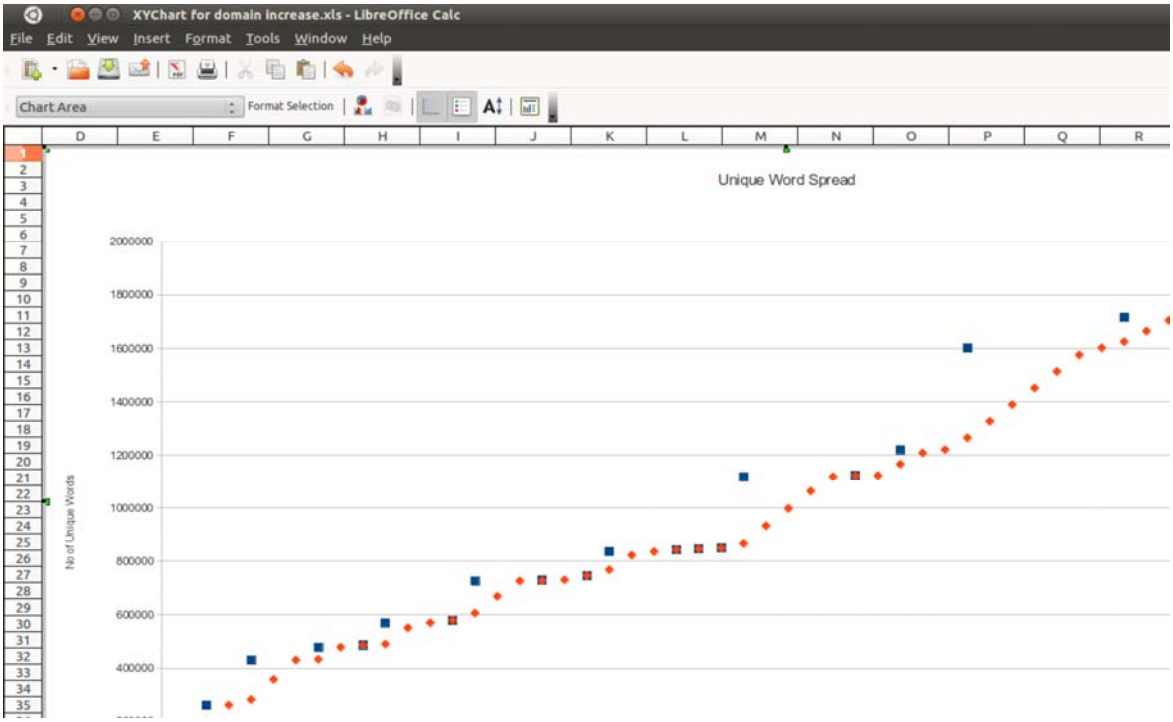
கையெழுத்து மற்றும் அச்சிடப்பட்ட பனுவல்களைக் கையாளும் ஐரோப்பிய மொழிகளுக்கான ஆவணக் கண்டறிகை முறைமைகள் (document recognition systems) கண்டறிதலிலும், உணர்பிழைகளைத் திருத்துவதிலும் மீப்பெரும் சொல்லகராதிகளைப் பயன்படுத்துகின்றன.

படம் 6, தமிழ் மொழியின் ஒரு முக்கியமான இயல்பைத் தெளிவாகக் காட்டுகிறது. ஒவ்வொரு ஆள்களமாகச் சேர்த்துக் கொண்டு செல்கையில், தனிச்சொற்களின் மொத்த எண்ணிக்கை, தவிட்டாமல், மென்மேலும் வளர்ந்து கொண்டே இருப்பதுதான் அது. ஆகவே, தமிழ் மொழியானது, ஒரு வரையறுக்க முடியாத தனிச்சொல் வங்கியை உடையது. அதனால், சொல் மட்டப் பேரகராதி (word level lexicon) கொண்டு பிழை திருத்துவது என்பது தமிழ் மொழிக்குச் சரிவராது. எங்கள் ஆய்வில், எமில்லெ (Emille) பனுவலிலிருந்து தேர்ந்தெடுக்கப்பட்ட 2.4 லட்சம் தனிச்சொற்கள் கொண்ட பேரகராதியில் நாங்கள் கையெழுத்து உணர்வியைச் சோதனை செய்ய உபயோகித்த சாதாரணமான 2000 வார்த்தைகளில் 1530 வார்த்தைகளே இருந்தன [26] என்பது குறிப்பிடக்கது. ஈரெழுத்துப் (bigram) படிமங்களால் மட்டுமே கையெழுத்து உணர்வியின் (online handwriting recognition system) பிழைகளைத் திருத்த இயலும் [27]. அப்படியும் பல ஜோடி வார்த்தைகள் (உ: அவன், அவள்; வருவான், வருவாள்) கடைசி எழுத்தில் மட்டும் வேறுபடும் என்பதால், அவற்றிற்கு ஈரெழுத்துப் படிமங்களும் தவறான சொற்களைக் கொடுக்க வாய்ப்பு உள்ளது. ஆக, அத்தகைய சூழ்நிலையில் நாம் உணர்வியின் எழுதப்பட்ட சொல்லிலிருந்து கீற்றக்குறியீடுகளுக்கு (symbols) துண்டுபடுத்தல் (segmentation) [28] மற்றும் உணர்தல் (recognition) ஆகிய பணிகளின் துல்லியத்தையே முயன்று அதிகரிக்க வேண்டியிருக்கிறது [29]. இந்தக் கருத்து, ஒளிவழி எழுத்துரு உணர்தல் (OCR) முறைமைகளுக்கும் [30, 31, 32] பொருந்தும்.

தமிழில் நீளமான சொற்கள் அதிகம்

தமிழில் ஒவ்வொரு வினைச்சொல்லுக்கும் குறைந்த பட்சமாக எண்பது உருபனியல் அமைப்புகள் (paradigms) உள்ளன. கூட்டு வினைச்சொற்களையும் கணக்கில் கொண்டால் இது இன்னும் பல மடங்காக உயரும் [22]. தமிழ் உருபனியல் மாற்றங்கள் எழுவாய், பயனிலை, செயப்படுபொருளைக் கண்டறிவதன் மூலம் பிரதிபெயர்களின் ஆதியைக் கண்டறிதலில் (pronominal resolution) பயன்படுமளவுக்கு பொருள் செறிந்தது [33]. இத்தகைய செறிவு, தெலுங்கு போன்ற மற்ற திராவிட மொழிகளிலும் உண்டு [34]. இச்செறிவே சொற்திருத்தி போன்ற முறைமைகளை

உருவாக்குவதைக் கடினமாக்குகின்றன [35]. இவ்வித உருபனியல் செறிவானது, ஒரு ஆங்கிலச் சொற்றொடரை மொழிபெயர்க்கையில் ஒரு வார்த்தையில் பொருள் தரும் வகையில் அமைந்துள்ளது [36]. உருபனியல் மாற்றச் செறிவினால் தமிழ் வார்த்தைகளின் சராசரி நீளம் அதிகமாக இருக்க வாய்ப்பு உள்ளது. இதனைச் சோதிப்பதற்காக வெவ்வேறு நீளமுள்ள தமிழ் தனி வார்த்தைகளின் புள்ளி விவரங்களைச் சேகரித்தோம். இவை அட்டவணை 6-ல் கொடுக்கப்பட்டுள்ளன. இங்கு நீளம் என்பது தமிழ்ச் சொல்லிலுள்ள உயிர்மெய் எழுத்துக்களின் எண்ணிக்கையே . உதாரணமாக, "கல்கி" என்ற சொல்லில் மூன்று எழுத்துக்கள் உள்ளன. ஆயின், ஒருங்குறி எண்ணிக்கை ஐந்தாகும் .



படம் 6. தமிழ் மொழியின் செறிவு : பனுவலில் ஒவ்வொரு ஆள்களமாகச் சேர்க்கும்போது (நீல நிறம்) தனிச் சொற்களின் எண்ணிக்கை தங்கு தடையின்றி வளர்தல். ஒவ்வொரு தடவையும் மற்றுமொரு லட்சம் பனுவல் சேர்க்கப்படும்பொழுது தனிச்சொற்களின் வளர்ச்சி (சிவப்பு நிறம்)

அட்டவணை 6: வெவ்வேறு நீளமுள்ள தமிழ் தனி வார்த்தைகளின் புள்ளி விவரங்கள் . மொத்த எண்ணிக்கையை ஒட்டி, இறங்கு வரிசைப்படி கொடுக்கப்பட்டுள்ளன .

எழுத்துக்கள் /சொல்	மொத்த எண்ணிக்கை	எழுத்துக்கள் /சொல்	மொத்த எண்ணிக்கை
10	135486	5	41230
11	130425	18	30471
9	128475	4	23679
12	122741	19	21826
8	115506	20	14660
13	107806	21	10195
7	90559	3	8465
14	90164	22	6818
15	72450	23	4512
6	64978	24	3035
16	55613	25	1924
17	41860	2	1833

முடிவுரை மற்றும் வருங்கால ஆய்வுப் பணிகள்

பை-கிராம் இடுகுறி அல்லது எழுத்து சார்ந்த புள்ளி விவரங்கள் வெவ்வேறு ஆள்களங்களுக்குக் குறிப்பிடத்தக்க அளவில் வேறுபடுகிறதா என்பதை உற்றறிதல் மூலம் பல பயனுறு தகவல்களைப் பெறலாம். உதாரணமாக, இப்புள்ளிவிவரங்கள் சங்கத்தமிழ்ப் பாடல்களுக்கும் நவீன தமிழ் தரவுகளுக்கும் குறிப்பிடத்தக்க விதத்தில் வேறுபட்ட தனிச்சிறப்பியல்புகளைக் கொண்டிருக்கும். மேலும் இலக்கண, சொற்பொருளியல் மற்றும் புணர்ச்சி விதிகளை வரையறுப்பதன் மூலம் உணர்தல் நெறிமுறைகளின் செயலாக்கம் வலுவடையலாம்.

நன்றி (Acknowledgment)

இந்தக் கட்டுரை யிலுள்ள கருத்துக்கள், இந்திய அரசின் இந்திய மொழிகளுக்கான தொழில் நுட்ப வளர்ச்சித் திட்டத்தின் (Technology Development for Indian Languages – TDIL) நிதி உதவியுடன் நாங்கள் மேற்கொண்ட ஆய்வுச் செயல் திட்டங்களில் பணி புரியும்போது பெற்ற அனுபவங்கள், எதிர்கொண்ட இடையூறுகள் முதலியவற்றால்

ஊக்குவிக்கப்பட்டன. அதற்காக, நாங்கள் இந்திய அரசின் தொலைத்தொடர்பு மற்றும் தகவல் தொழில் நுட்பத் துறைக்குக் கடமைப் பட்டுள்ளோம். மேலும், முதன்மையாக, எங்கள் திருக்குரல் ஒளிவழி எழுத்துரு உணர்தல் முறைமையைப் பயன்படுத்தும் பலரும் கொடுத்த தமிழ் உரைகளிலிருந்தே இந்தக் கட்டுரைக்கான பனுவல் தயாரிக்கப்பட்டது. அதற்காக, அவர்கள் அனைவருக்கும் எங்கள் நன்றி. மற்றும் பேராசிரியர் தெய்வசுந்தரம் அவர்களும், முனைவர். வாசு ரெங்கநாதன் அவர்களும் இந்தக்கட்டுரையை நாங்கள் வழங்கும் முறையின் தரத்தை உயர்த்த ஆலோசனைகள் வழங்கினார்கள். அவர்களுக்கும் எங்கள் நன்றிகள் பல.

மேற்கோள்கள் (References)

1. Shiva Kumar H R, Ashwini J K, Rajaram B S R and A G Ramakrishnan, “MILE TTS for Tamil and Kannada for blizzard challenge 2013,” Proc. of Blizzard Challenge Workshop, Barcelona, Spain, September 3rd 2013.
2. Web demo of Thirukkural Tamil text to speech conversion system. http://mile.ee.iisc.ernet.in:8080/tts_demo/
3. G. L. Jayavardhana Rama, A. G. Ramakrishnan, R. Muralishankar and P. Prathibha, “A Complete Text-to-Speech Synthesis System in Tamil,” Proc. IEEE 2002 Workshop Speech Synthesis, Santa Monica, CA USA, Sep. 11-13, 2002, pp. 191-194.
4. Pertti Palo, “A Review of Articulatory Speech Synthesis,” Master’s Thesis, Helsinki University of Technology, June 2006.
5. K. Partha Sarathy and A. G. Ramakrishnan, “A Research bed for unit selection based text to speech synthesis,” Proc. IEEE Workshop on spoken language technology (SLT 08), Dec. 15-18, 2008, Goa, India.
6. ஆ. க. ராமகிருஷ்ணன், ப. அருள்மொழி, “பல்வகை உணர்விகளில் மொழி மாதிரியங்களின் பயன்பாடு : ஒரு புதிய கண்ணோட்டம்,” 12-வது உலகத் தமிழ் இணைய மாநாட்டு மலர், கோலாலம்பூர், மலேசியா, ஆகஸ்ட் 15-18, 2013.
7. K. Suresh and A. G. Ramakrishnan, “A DCT based approach to Estimation of Pitch,” Proc. Intern. Conf. Multimedia Processing and Systems, Chennai, Aug. 13-15, 2000, pp. 54-57.
8. R. Murali Shankar and A. G. Ramakrishnan, “Robust Pitch detection using DCT based Spectral Autocorrelation,” Proc. Intern. Conf. Multimedia Processing and Systems, Chennai, Aug. 13-15, 2000, pp. 129-132.
9. A. P. Prathosh, T. V. Ananthapadmanabha, and A. G. Ramakrishnan, “Epoch extraction based on integrated linear prediction residual using plosion index,” IEEE Transactions on Audio, Speech and Language Processing, 2013, Vol. 21, Issue 12, pp. 2471-2480.
10. R Muralishankar, A. G. Ramakrishnan and P Prathibha, “Modification of Pitch using DCT in the Source Domain,” Speech Communication, 2004, Vol. 42/2, pp. 143-154.

11. R. Muralishankar, A. Vijay Krishna and A. G. Ramakrishnan, “Subspace based Vowel Consonant Segmentation,” Proc. IEEE Workshop on Statistical Signal Processing, Sept 28 - Oct 1, 2003, St. Louis, Missouri, pp. 589- 592.
12. A. G. Ramakrishnan, Lakshmish N. Kaushik and Laxmi Narayana M, “Natural Language Processing for Tamil TTS,” Proc. 3rd Language and Technology Conference, Poznan, Poland, Oct 5-7, 2007, pp. 192-196.
13. A. G. Ramakrishnan and Laxmi Narayana M, “Grapheme to Phoneme Conversion for Tamil Speech Synthesis,” Proc. Workshop in Image and Signal Processing (WISP-2007), IIT Guwahati, Dec 28-29, 2007, pp. 96-99.
14. Vikram Ramesh Lakkavalli, Arulmozhi P and A G Ramakrishnan, “Continuity Metric for Unit Selection based Text-to-Speech Synthesis,” IEEE International Conference On Signal Processing & Communications (SPCOM 2010), 2010, Bangalore, India.
15. Sreekanth Majji and A G Ramakrishnan, “Festival Based Maiden TTS System for Tamil Language,” Proc. 3rd Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, Poznan, Poland, Oct 5-7, 2007, pp. 187-191.
16. H R Shiva Kumar, Abhinava Shivakumar, Akshay Rao, S Arun, A G Ramakrishnan, “Panmozhi Vaayil - A Multilingual Indic Keyboard Interface for Business and Personal Use,” Proc. Information Systems for Indian Languages, Patiala, 2011.
17. G. L. Jayavardhana Rama, A. G. Ramakrishnan, M. Vijay Venkatesh and R. Muralishankar, “Thirukkural - a text-to-speech synthesis system,” Proc. Tamil Internet 2001, Kuala Lumpur, August 26-28, 2001, pp. 92-97.
18. Kishore Prahallad, Anandaswarup Vadapalli, Naresh Elluru, Gautam Mantena, Bhargav Pulugundla, Peri Bhaskararao, Hema A. Murthy, Simon King, Vasilis Karaiskos and Alan W Black, "The Blizzard Challenge 2013 -- Indian Language Task", Proc. of Blizzard Challenge Workshop, Barcelona, Spain, September 3rd 2013.
19. R. Muralishankar and A. G. Ramakrishnani, “Human touch to Tamil Synthesizer,” Proc. Tamil Internet 2001, Kuala Lumpur, August 26-28, 2001, pp. 103-109.
20. A G Ramakrishnan and Lakshmi Chithambaran, “Modeling basic emotions for Tamil speech synthesis,” Proc. 12-th International Tamil Internet Conf., Kuala Lumpur, Malaysia, Aug. 15-18, 2013.
21. K. Partha Sarathy and A. G. Ramakrishnan, “Text to speech synthesis system for mobile applications,” Proc. Workshop in Image and Signal Processing (WISP-2007), IIT Guwahati, Dec 28-29 2007, pp. 74-77.
22. Rajendran, S., Viswanathan, S. Ramesh Kumar, “Computational morphology of Tamil verbal complex,” Language in India, Vol. 3 : 4 April 2003.
23. Annamalai, E. Dynamics of Verbal Extension in Tamil. Thiruvananthapuram: Dravidian Linguistics Association, 1985.
24. Richard Lederer. A celebration of English, good grammar, and wordplay. Marion Street Press, 2012.
25. David Crystal, "The Stories of English", Overlook TP Publishers, 2004. ISBN:978-1-58567-719-1
26. Suresh Sundaram, Bhargava Urala and A. G. Ramakrishnan, “Language models for online handwritten Tamil word recognition,” Proc. Workshop on Document Analysis and Recognition (DAR 2012), 16 December 2012, IIT Bombay, Mumbai, India.
27. Suresh Sundaram and A. G. Ramakrishnan, “Bigram language models and reevaluation strategy for improved recognition of online handwritten Tamil words,”

revised manuscript under review, ACM Transactions on Asian Language Information Processing (TALIP), 2013.
28. Suresh Sundaram and A. G. Ramakrishnan, “Attention-feedback based robust segmentation of online handwritten isolated Tamil words,” ACM Transactions on Asian Language Information Processing (TALIP), Vol. 12 (1), March 2013, Article No. 4.
29. Suresh Sundaram and A. G. Ramakrishnan, “Performance enhancement of online handwritten Tamil symbol recognition with reevaluation techniques,” Pattern Analysis and Applications, Dec. 2013.
30. K. G. Aparna and A. G. Ramakrishnan, “A complete Tamil Optical Character Recognition System,” Proc. Fifth IAPR Workshop on Document Analysis Systems DAS-02, Princeton, NJ, August 19-21, 2002, pp. 53-57.
31. A. G. Ramakrishnan and Kaushik Mahata, “A Complete OCR for Printed Tamil Text,” Proc. Tamil Internet 2000, Singapore, July 22-24, 2000, pp. 165-170.
32. D. Dhanya and A.G.Ramakrishnan, “Simultaneous recognition of Tamil and Roman scripts,” Proc. Tamil Internet 2001, Kuala Lumpur, August 26-28, 2001, pp. 64-68.
33. Murthy, Kavi Narayana, L. Sobha, and B. Muthukumari, "Pronominal resolution in Tamil using machine learning," Proc. First Intern. Workshop on Anaphora Resolution (WAR-I), pp. 39-50. 2007.
34. K. Narayana Murthy, "Parsing Telugu in the UCSG Formalism," Proc. Indian Congress on Knowledge and Language, vol 2, Jan. 1996, pp 1-16, Central Institute of Indian Languages, Mysore.
35. Ranjani Parthasarathy and Geetha T.V., Morphological Analyzer for Tamil, TI 2001, Chennai.
36. Vu, Ngoc Thang, and Tanja Schultz, "Initial experiments with Tamil LVCSR," Proc. Intern. Conf. Asian Language Processing (IALP), Hanoi, Vietnam. pp. 81-84, IEEE, 2012.