Significance of Paralinguistic Cues in the Synthesis of Mathematical Equations

Venkatesh Potluri, SaiKrishna Rallabandi, Priyanka Srivastava, Kishore Prahallad
International Institute of Information Technology - Hyderabad
{venkatesh.potluri, saikrishna.rallabandi}@research.iiit.ac.in
{priyanka.srivastava, kishore}@iiit.ac.in

Abstract

Text to speech (TTS) systems hold promise as an information access tool for literate and illiterate including visually challenged. Current TTS systems can convert a typical text into a natural sounding speech. However, auditory rendering of mathematical content, specifically equation reading is not a trivial task. Mathematical equations have to be read so that appropriate bracketing such as parentheses, superscripts and subscripts are conveyed to the listener in an accurate way. Earlier works have attempted to use pauses as acoustic cues to indicate some of the semantics associated with the mathematical symbols. In this paper, we first analyse the acoustic cues which human-beings employ while speaking the mathematical content to (visually challenged) listeners and then propose four techniques which render the observed patterns in a text-to-speech system. The evaluation considered eight aspects such as listening effort, content familiarity, accentuation, intonation, etc. Our objective metrics show that a combination of the proposed techniques could render the mathematical equations using a TTS system as good as that of a human-being.

1 Introduction

Mathematical equations comprise of different types of visual cues to convey their semantic meaning. Some of these visual cues are superscripts, subscripts, parentheses,etc. Despite advances in screen reading and text to speech technologies, the problem of speaking complex math remains majorly unsolved. Speaking the equation just as any other string of text, a line, or a sentence will not suffice to effectively render mathematics in speech. For instance, $e^{x+1} - 1$ denotes that the value "e" should be multiplied "x+1" times before subtracting 1 from it. However, when it is rendered in speech like a general string, it is difficult to identify the portion of the equation in the superscript and the remainder of it after the superscript. To effectively resolve such ambiguities and identify such demarcations in mathematical content, information presented through visual cues such as spatialisation must be mapped to their auditory equivalent. Mathematics, in its visual form, gives the reader a very high level granularity in perceiving the equation. Mathematical equations, when presented in audio must be able to match the advantage in granularity provided in visual representation of mathematics. The typical issues in audio rendering of mathematical equations include quantification, superscripting and subscripting, and fractions.

1.1 Quantification

Most of mathematical equations contain expressions in parentheses. For instance, considering the equation $(A+B)*(C+D)+E$, it may seem that the equation can just be treated as a general string of text while speaking. However, this will create a confusion in the listener, as there are two ways of expressing.

- "left parenthesis A plus B right parenthesis times left parenthesis C plus D right parenthesis plus E"
- "A plus B times C plus D plus E".

In the former case, the listener will have to keep a track of all the parentheses when he or she listens to the equation. This becomes a hectic task for bigger equations and also results in deviating the listener's attention from concentrating on the actual contents of the equation. On the other hand,

D S Sharma, R Sangal and J D Pawar. Proc. of the 11th Intl. Conference on Natural Language Processing, pages 395–402,
Goa, India. December 2014.

in the latter case, the listener gets an ambiguous representation of the equation. The spoken form of the equation should have additional information to the equation to solve this ambiguity.

1.2 Superscript and subscript

Today's screen readers and TTS engines do not effectively convey the equations with superscript and subscript content. They often do not speak out the parts of the equation contained in the superscript and subscript. They often speak out such content continuously, with the rest of the equation. For instance, let us say the expression is E^X. With the currently available technologies, the expression may be rendered as "EX". This does not give the listener the information that X is in the superscript and the listener may understand the expression as $E * X$. In expressions where there are at least 2 variables that cause a phonetic sound when spoken together, the general TTS may treat the expression as a complete word. consider the expression A^B. The TTS may speak it as "ab". In case of numbers, say we have an expression 5^{2^5}, the TTS reads it as "five hundred twenty five" or "five two five". We come across the same issues while trying to render subscript text. If a human speaks the expression, he may not make such mistakes. The challenge to the human speaker lies in effectively conveying the spatial orientation of the different parts of the equation. That is, the equation, presented in audio must give the listener a clear picture of what content is in the superscript and the subscript. The listener must also be able to observe the end of the super script or subscript part of a mathematical expression. The listener should understand that any thing that he listens to after the end is in the baseline or the general part of the equation, unless specified. To overcome this challenge, the spoken form of an equation should provide the listener with different cues for superscript and subscript content.

1.3 Fractions

Fractions, like the other mathematical concepts discussed above can not be treated like a general string of text. The key information that has to be conveyed to the listener in addition to the contents of the fraction is the beginning of the fraction, the content of the fraction in numerator and denominator and the end of the fraction. The audio equivalent of the equation should effectively be able to convey nested fractions in addition to the regular fractions to the listener.

There have been several attempts to present mathematical content through alternative modes to vision. Efforts have been made to formulate standards for presenting math through Braille and speech. Nemeth Code(Nemeth et al., 1973) is a special type of Braille used for math and science notations. With Nemeth Code, one can render all mathematical and technical documents into six-dot Braille. This code could also be used to speak mathematical content. Dr T.V Raman has developed an audio system for technical readings (ASTER)(Raman, 1998). ASTER is a computing system for producing audio renderings of electronic documents. The present implementation works with documents written in the TEX family of markup languages: TEX, LaTeX and AMS-TEX. A more recent attempt has been made by a company called design science. They developed an internet explorer plugin called MathPlayer that displays and speaks out mathematical content marked up in MathML(Soiffer, 2005). There have been attempts to form a set of guidelines to effectively speak mathematics in audio. The handbook for spoken mathematics (Chang et al., 1983) gives an account of such an attempt. An article on how to speak math also describes the challenges in speaking mathematics to and by a computer (Fateman, 1998). The ChromeVox project (Raman et al., 2012) is a screen reader built for Google Chrome browser and the Chrome OS. It has basic support for mathematical expressions encoded using the MathML language on web pages. The expressions are verbally presented during normal text navigation. The screen reader announces that the spoken text is a mathematical expression and it can further be explored. Navigation support is based on the MathML tree.

Earlier works discussed so far, have not effectively used paralinguistic cues and variations in the equation. However, humans use a lot of cues when reading out a mathematical equation which helps in understanding the semantics of it. Usage of the cues similar to the humans would result in more effective rendering of the equations.

The objective of this paper is to analyse the way these visual cues are presented in an auditory format by human speakers who are well acquainted with speaking the mathematical content especially to visually challenged individuals. A subjective

and objective analysis is performed on the equations recorded by the speakers. Based on this analysis, we make an attempt to form specific rules to map the visual cues to their auditory equivalents to programatically and unambiguously render the mathematical content in audio using a text-to-speech system.

Section 2 discusses the basis for the study. Section 3 has the inferences drawn form the initial listening tests . Section 4 discusses the proposed ideas. Section 5 presents the analysis of the qualitative study performed.

2 Cues in spoken equations

Our study is based on the preposition that treating a mathematical expression as a regular English sentence while speaking is not an effective way to present mathematical content in an auditory form. In order to test this observation, we asked a set of 15 people to rate mathematical equations spoken by a traditional TTS system. Then we conducted the same experiment on spoken equations (i.e., equations spoken by a human-being). The details of the listening tests are as follows.

2.1 Procedure for the listening tests

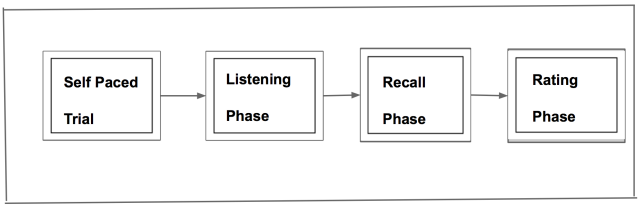


Fig: Evaluation procedure

A set of 15 participants were made to listen to the recorded equations. Each participant was made to listen to the equations using headphones and the responses were recorded. The listening test was self paced and also the users were informed that they were free to listen to the equation any number of times till they felt comfortable that they could recall the equation. Similarly, the same participants were also made to listen to speech of mathematical equations generated by a TTS system. The participant will have to reproduce the equation he/she listens to. In addition to reproducing the equation, the participant will have to evaluate the spoken equation based on eight other parameters, i.e., perform objective analysis. We arrived at these parameters partly by following the listening test procedures followed in the Blizzard challenges (Hinterleitner et al., 2011) and our own analysis.

Table 1: Evaluation of Spoken Math vs TTS

Parameter	Spoken	Synthesized (Current TTS)
Listening Effort	2.5	4.4
Content Familiarity	2.7	2.7
Effectiveness of additional cues	3.2	1.2
Accentuation	4.3	2.5
Intonation	4.26	1.6
Pauses	3.1	2.15
Number of repetitions (Mode)	2	4
Mean Opinion Score	4.42	1.89

2.2 Selection of the equations

Selection of suitable equations is a critical component to analyse the auditory presentation of mathematical content. We hand picked a few equations which had variations in number of variables, number of sub expressions and length of the equation. The equations can be found in appendix A. Each of the equations is semantically unrelated, that is, the equations have mathematical content but the listener may not have come across the exact same equation prior to listening to them from our recordings . The reason behind choosing the equations in such a way is to ensure that the listener's prior knowledge does not influence the ability to recall the equation. If the listener is able to recall the equation even before he or she listens to it completely, the listener is benefitting from memory, not the spoken equation.

2.3 Parameters for objective analysis

On a scale of 1 to 5, the participants were asked to evaluate the spoken equations on the following parameters.

- Listening effort (1 = low, 5 = high)
- Intonation (1 = ineffective and 5 = very effective)
- Acceptance (1 = poor, 5 = good).
- Speech pauses (1= not noticeable and 5 = very prominent)

- Accentuation (1 = poor and 5 = very prominent).
- Content familiarity (1 = totally new concept and 5 = very familiar). Here 1 indicates that the user is not acquainted to the terminology used in the equation. In this case, the participants' response for that particular equation can not be considered completely as he may have entered a wrong response due to the lack of domain knowledge, not due to the lack of understanding of the audio.
- Effectiveness of additional cues such as sounds, pitch and rate variations, change in direction, etc. (1 = hardly noticeable and 5 = very helpful).
- Number of repetitions of each equation.

3 Inferences from the listening tests

The results of this experiment, shown in the Table 1 indicate that the equations are not intelligible enough if it is spoken as a plain text using a text-to-speech system. The mean opinion scores of spoken equations indicate a human-being use several acoustic cues to manifest the semantics of the mathematical symbols in audio mode. It was noticed that the trained speakers brought certain variations in their speech while speaking specific aspects of the mathematical expression. The variations are noticed in pauses and pitch variations (intonation). A careful analysis revealed that the acoustic variations were introduced by the speakers to unambiguously speak 1) quantification, 2) superscripting and subscripting and 3) handling fractions in mathematical equations.

Based on the feedback received from participants, we can infer that the use of these additional cues can effectively and unambiguously present mathematical content in audio. The question is how to introduce such cues to synthesise a mathematical equation using a text-to-speech system.

4 Proposed techniques

With the advent of languages like MathML, it is possible to programatically identify different attributes and visual cues of a mathematical expression. This possibility can in turn be leveraged to make some modifications while generating speech for mathematical content. We propose four techniques that could enhance the way mathematical content is rendered in audio.

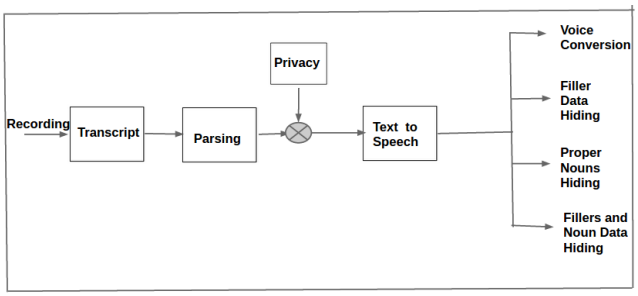


Fig: Overall framework for the proposed techniques

An example depicting the workflow of the entire algorithm is shown in the Figure 1. For the sake of illustration, a simple expression , $(X+Y)^{4-2}$ was taken :

The Equation was first converted into the Math Markup Language format. We chose "Presentation" Markup style to represent the equations. It is then text processed to identify and segregate the different terms occurring in the equation. The following terms have been segregated.

- Subscripts and superscripts
- Fractions
- Square root terms
- Overscripts and underscript

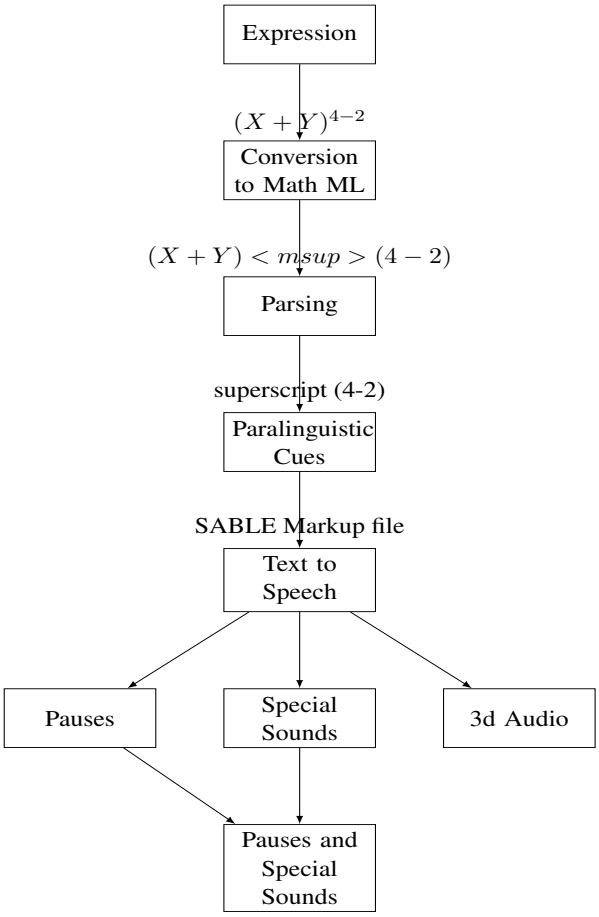
The MathML representation is processed to convert it into natural language and the acoustic cues such as pauses, intonation are incorporated to generate a file in the SABLE[1] markup language. The SABLE file is input to the speech synthesis system which generates the audio form of the equation with specified pauses and intonation. We have generated the audio files using the Festival Speech Synthesis System(Black et al., 2002). Sections 4.1 through 4.4 discuss each of the four proposed techniques.

4.1 Technique 1 : Rendering equations with pauses and special sounds

In visual communication, icons and symbols are used as indications for some types of information. In the context of mathematical expressions, the user can perceive the type of elements (

[1]SABLE is mark up language due to collaboration between Sun, AT&T, Bell Labs, Edinburgh and CMU to devise a standard cross synthesizer standard mark up language. The language is XML-based and allows users to add addition controlling commands in text to affect the output. An implementation exists in Festival speech synthesis system.

Figure 1: Example Synthesis using a simple expression



superscripts, subscripts, etc) by getting a glance at the equation. A person has the advantage of perceiving a lot of information of the equation even before looking at the actual contents of the equation. This technique attempts to present the equation in a manner that a person gets a similar advantage when he listens to it.

In this concept, we made use of special sounds or ear cons while presenting the equations. However, replacing speech with sounds alone is not the most effective way to tackle the problem of presenting mathematic equations in audio. We made use of paralinguistic cues including, but not limited to sounds.

The cues presented in this method include:

- **Pauses** to convey certain parts of an equation. These pauses are mainly used to separate the parts of mathematical expressions. Consider $(A+B)^2$ and $(A+B^2)+1$. It would sound more natural and intuitive if the expressions are spoken as "the quantity A + B pause superscript 2 " and "the quantity A + B superscript 2 pause + 1" .

- **Sounds** to indicate certain symbols and mathematical operations. Sounds are used to indicate superscripts, subscripts, roots, under scripts, over scripts and under script-over script combination.

Table 2: Pitch and rate variations

Term	Pitch variation	Rate variation
Superscript	50	20
Subscript	-50	-20
Fraction	25	-25
Underscript	-60	-25
Overscript	60	25

We chose the sounds(such as the sound "ding") such that would be pleasant to the ear and that are passively noticed by a listener so as not to distract too much, at the same time, are loud enough not to go unnoticed. The sounds show a transition from high to low and low to high when there is a subscript and superscript respectively. Any other type of sounds and their variations could also be applied in this technique.

4.2 Technique 2 : Rendering equations with pitch and rate variations

Screen Reader users are familiar to pitch changes. Generally, a high pitch is used to denote capitals and a low pitch is used to denote tool tip messages. On observing the human recorded equations explained in Section 2, we observed that speakers tend to modulate the pitch as they read aloud certain parts of a mathematical expression. It has been observed that certain parts of a mathematical expression are spoken at a faster rate to indicate that it is a sub expression and to isolate it from the rest of the expression.

In this technique, we use pitch and rate changes to denote the presence of certain mathematical attributes. The pitch and rate increase while speaking out the superscript text and decrease while speaking the subscript text. A similar method can be employed to properly render fractions. The numerator is spoken in a higher pitch and the denominator is spoken in a lower pitch. Similarly, quantities in a root are spoken at a faster rate. Table

Table 3: Evaluation of the proposed techniques

Parameter	Technique#1	Technique#2	Technique#3	Technique#4	
Intonation Variation	2.3	**4.7**	4.32	4.68	
Pitch Variation	1.4	4.43	**4.82**	4.36	
Pauses	**4.15**	3.7	3.7	3.87	
Listening Effort	3.5	**2.3**	2.64	2.47	
Content Familiarity	2.7	2.7	2.7	2.7	
Effectiveness of additional cues	1.82	4.32	**4.37**	4.23	
Accentuation	**3.47**	2.3	3.2	3.6	
Number of repititions(Mode)	3	2	2	2	
Mean Opinion Score	2.27	4.37	**4.62**	4.35	

2 shows the pitch and rate variation(in percentage) that are applied to the Mathematical equation. The variation is with respect to the base pitch and rate of the TTS.

4.3 Technique 3: Rendering equations with audio spatialisation

In this technique, we made an attempt to draw a closer analogy to the spatial positioning of various variables and numbers of a mathematical equation. The listener can be given the illusion that the superscript part of the math expression is spoken from above his head and the rest at the usual level using the Head Related Transfer Function (HRTF) (Geronazzo et al., 2011). Table 4 shows the sets of angles chosen for the different parts of the equation such as superscript, etc.

Table 4: Sets of HRTF angles for audio spatialisation

Term	Elevation Angle	Azimuth Angle
Superscript	90	30
Subscript	-90	30
Fraction	270	45
Underscript	-90	45
Overscript	90	30

We identify the portions of a mathematical expression that require modification in spatial orientation of sound. Based on the attribute, we apply the HRTF function with the required angles.

4.4 Technique 4 : Rendering equations with pitch variations and special tones

In this technique, we render the equations in audio by varying the pitch, adding pauses, emphasising the speech and adding sounds at required parts of a mathematical expression. As explained in 4.3, we can make pitch and rate manipulation while rendering superscripts, subscripts, fractions, under scripts and over scripts. In addition to the variations in speech, we have also added sounds to indicate the listener before hand that he must expect one of the above mentioned variations (superscripts, subscripts, etc). The sounds used here are the same as the ones mentioned in section 4.2. The Pitch and rate variations that are introduced are the same as the percentage values given in table 2.

5 Analysis of the listening test

A system was built to render mathematical expressions implementing each of the proposed ideas. An experiment procedure similar to the one explained in Section 2 was followed. 30 participants were made to participate in the experiment. The table contains the normalised scores(1 to 5) calculated over the responses for the equations. The number of repetitions of the equation has the mode value(most occurring value).

On analysing the experiment as described in Section 2, it is observed that the participants are able to understand the human spoken equations. More over, it can be clearly understood that generating spoken forms of mathematical equations without making any enhancements is not capable of rendering math effectively. It can also be inferred that making use of just a few paralinguistic cues, sounds and pauses as explained in section 4.1 will not suffice either. The pitch and rate changes while rendering certain parts of the mathematical expressions have proven to be helpful to the participants in comprehending the expression. In the method described in section 4.3, the lis-

tener has been able to draw an analogy to the print form of mathematics. It has been observed that the method explained in section 4.1 did not prove to be helpful to the listeners. However, from the table 3 and the values corresponding to the technique explained in section 4.4, it is evident that use of cues (pauses and rate variations) in addition to special sounds can be significantly effective in helping a listener.(see the demonstration of the listening test on the webpage associated to this paper: http://goo.gl/FLTIOv).

6 Conclusion

From the analysis and the proposed ideas, we can say that there is a possibility to unambiguously render mathematics in audio. With the increase in voice driven interfaces and information access through audio, rendering mathematical content in audio could also help more effectively present such content in these interfaces. Personal assistance or any other voice driven UIs can more effectively render mathematical content to the listener. In addition to this, effectively rendering mathematical content in audio can be of a great advantage for people with print disabilities including, but not limited to vision impairment, dyslexia and cognitive impairment. With currently available assistive technology, understanding mathematical content is very difficult and almost impossible. the ideas explained in sections 4.1 to 4.3 improve the scenario of understanding mathematical content through a non visual input mode. as explained in section 4.4, There is also a chance that a combination of the proposed ideas are more effective than each of the ideas alone.

Acknowledgements

We thank Prof Peri Bhaskararao for his contributions in our initial discussions related to this research effort.

References

Alan W Black, Paul Taylor, Richard Caley, and Rob Clark. 2002. The festival speech synthesis system. *University of Edinburgh*, 1.

Larry A Chang, CM White, and L Abrahamson. 1983. Handbook for spoken mathematics. *Lawrence Livermore National Laboratory*.

Richard Fateman. 1998. How can we speak math. *Journal of Symbolic Computation*, 25(2).

Michele Geronazzo, Simone Spagnol, and Federico Avanzini. 2011. A head-related transfer function model for real-time customized 3-d sound rendering. In *Signal-Image Technology and Internet-Based Systems (SITIS), 2011 Seventh International Conference on*, pages 174–179. IEEE.

Florian Hinterleitner, Georgina Neitzel, Sebastian Möller, and Christoph Norrenbrock. 2011. An evaluation protocol for the subjective assessment of text-to-speech in audiobook reading tasks. In *Proceedings of the Blizzard challenge workshop, Florence, Italy*. Citeseer.

Abraham Nemeth, National Braille Association, et al. 1973. *The Nemeth Braille Code for mathematics and science notation*. American Print. House for the Blind.

TV Raman, Charles L Chen, and Dominic Mazzoni. 2012. Rachel shearer, chaitanya gharpure, james deboer, david tseng google inc 1600 amphitheatre parkway.

TV Raman. 1998. *Audio system for technical readings*. Springer.

Neil Soiffer. 2005. Mathplayer: web-based math accessibility. In *Proceedings of the 7th international ACM SIGACCESS conference on Computers and accessibility*, pages 204–205. ACM.

A Equations recorded by Human voice

$$X + Y = z \tag{1}$$

$$\frac{X+Y}{K} = \alpha \tag{2}$$

$$(X+Y)^{P+Q} = X^{P*Q} + Y^{P} * Q - P + \frac{Q}{Y} - \frac{P}{Q-X} \tag{3}$$

$$\frac{(P+X)*(Q-Y)}{(X+Y)^{K}} = \frac{P}{X+K} - Q*(\frac{K^{x}}{Y-P}) \tag{4}$$

$$(X+Y)^{K} = 3*X^{K} + 4*X^{y} - 5Y^{K+X} \tag{5}$$

$$(X+Y)^{P+Q} = X^{P*Q} + Y^{P} * Q - P + \frac{Q}{Y} - \frac{P}{Q-X} \tag{6}$$

$$\frac{(P+X)*(Q-Y)}{(X+Y)^K} = \frac{P}{X+K} - Q*(\frac{K^x}{Y-P}) \tag{7}$$

$$\frac{X+Y}{K} = \alpha \tag{8}$$

$$(X+Y)^K = 3*X^K + 4*X^y - 5Y^{K+X} \tag{9}$$

B Equations for testing the Systems

$$1+2+3-5+4+2+3 = (3+2)*(1+1) \tag{10}$$

$$\lim_{x \to +\infty} \frac{3x^2+7x^3}{x^2+5x^4} = 3. \tag{11}$$

$$\frac{\partial}{\partial x} x^2 y = 2xy \tag{12}$$

$$\frac{\partial u}{\partial t} = h^2 - E^{n+1} - 1 \tag{13}$$

$$\int_0^R \frac{2x\,dx}{1+x^2} = \log(1+R^2) \tag{14}$$

$$\int_0^{+\infty} x^n e^{-x}\,dx = n!. \tag{15}$$

$$(P+Q)^K + R = P^K * Q + Q^K * P + R^{P*Q} * K + \frac{P^Q * K + 1}{R} \tag{16}$$

$$(P+Q)*(R+K) = (P+R)^Q - (K+R^Q) + \frac{R+Q^K}{(R+Q)^K+1} \tag{17}$$

$$(P+Q)*(R+K) = (P+R)^Q - (K+R^Q) + \frac{R+Q^K}{(R+Q)^K+1} \tag{18}$$

$$\frac{X_1^K + X_2^K}{P_3^X * 5_4^x} + E^X = e^{\frac{X_{K+1}+X_{K+2}}{(X+Y)}} \tag{19}$$

$$\sqrt[P+Q]{A + K^P + A^{K+P}} = \frac{(K+P)(K-P)}{K*(P+K)} \tag{20}$$

$$\sum_{i=1}^{\infty} \frac{1}{i^2} + 5i + \sqrt[3]{i+1} = \frac{\pi^2 + 4\pi^3 + \sqrt[\pi+i]{9*\pi}}{6} \tag{21}$$

$$(\frac{X+Y}{K}+1)^3 = \sqrt[3]{X} + \sqrt[3]{Y} + (X*Y)/3 + \frac{X+Y}{3+K} + 3 \tag{22}$$